

国际信息工程先进技术译丛

CRC Press
Taylor & Francis Group

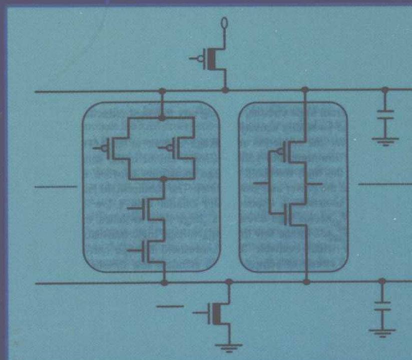
微电子技术原理、设计与应用（原书第2版）

Microelectronics (Second Edition)

（美） Jerry C. Whitaker 著

刘 波 曲新波 郭 飞 译





机械工业出版社
CHINA MACHINE PRESS

国际信息工程先进技术译丛

微电子技术原理、设计与应用（原书第2版）

（美）Jerry C. Whitaker　著
刘　波　曲新波　郭　飞　译

机械工业出版社

本书详细阐述了微电子技术的原理，并以技术原理与系统案例为线索，全面系统地介绍了特定微电子技术的各个要点和各种应用系统的设计与实现原理。同时，还分析了特定微电子技术在推动下一代微电子技术发展过程中所占的重要地位。

本书材料丰富、内容翔实、指导性强，可作为理工科院校电子、信息及相关专业的教材，也可作为电子技术人员和工程人员的技术参考书。

图书在版编目（CIP）数据

微电子技术原理、设计与应用（原书第 2 版）/（美）惠特克（Whitaker, Jerry.C.）著；刘波，曲新波，郭飞译. —北京：机械工业出版社，2008.5
（国际信息工程先进技术译丛）
ISBN 978-7-111-23879-9

Ⅰ. 微…　Ⅱ. ①惠…②刘…③曲…④郭…　Ⅲ. 微电子技术　Ⅳ. TN4

中国版本图书馆 CIP 数据核字（2008）第 050433 号

机械工业出版社（北京市百万庄大街 22 号　邮政编码 100037）
策划编辑：张俊红　责任编辑：朱　林　责任校对：申春香
封面设计：马精明　责任印制：杨　曦
北京外文印刷厂印刷
2008 年 6 月第 1 版第 1 次印刷
169mm×239mm·33 印张·643 千字
0001—4000 册
标准书号：ISBN 978-7-111-23879-9
定价：68.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
销售服务热线电话：（010）68326294
购书热线电话：（010）88379639　88379641　88379643
编辑热线电话：（010）88379764
封面无防伪标均为盗版

译 者 序

纵观人类社会发展的历史，一切生产方式和生活方式的重大变革都是由于新的科学发现和新技术的产生而引发的，科学技术作为变革的力量，推动着人类社会向前发展。从50多年前晶体管的发明到目前微电子技术成为整个信息社会的基础和核心的发展历史充分证明了"科学技术是第一生产力"。我们当前面临的信息革命是以数字化和网络化为特征——数字化大大改善了人们对信息的利用，更好地满足了人们对信息的需求；而网络化则使人们更为方便地交换信息，使整个世界"变得很小"。以数字化和网络化为特征的信息技术同一般技术不同，它具有极强的渗透性和基础性，它可以渗透和改造各种产业和行业，改变着人类的生产和生活方式，改变着经济形态和社会、政治、文化等各个领域。而它的基础之一就是微电子技术。可以毫不夸张地说，没有微电子技术的进步，就没有今天信息技术的蓬勃发展，微电子已经成为整个信息社会发展的基石。

微电子技术的发展历史实际上就是不断创新的过程，这里的创新包括原始创新、技术创新和应用创新等。晶体管的发明并不是一个孤立的精心设计的实验，而是一系列固体物理、半导体物理、材料科学等取得重大突破后的必然结果。1947年发明点接触型晶体管，1948年发明结型场效应晶体管以及以后的硅平面工艺、集成电路、CMOS技术、半导体随机存储器、CPU、非挥发存储器等微电子领域的重大发明也都是一系列创新成果的体现。同时，每一项重大发明又都开拓出一个新的领域，带来了新的巨大市场，对我们的生产和生活方式产生了重大的影响。也正是由于微电子技术领域的不断创新，才能使微电子元器件能够以每三年集成度翻两番、特征尺寸缩小一倍的速度持续发展几十年。在微电子技术发展的前50年，创新起到了决定性的作用，而今后微电子技术的发展仍将依赖于一系列创新性成果的出现。目前微电子技术已经发展到了一个很关键的时期，21世纪上半叶，也就是今后50年微电子技术的发展趋势和主要的创新领域主要有以下4个方面：以硅基CMOS电路为主流工艺；系统芯片（System on Chip, SoC）为发展重点；量子电子元器件和以分子（原子）自组装技术为基础的纳米电子学；与其他学科的结合诞生新的技术增长点，如MEMS（Micro Electro-Mechanical System，微机电系统），DNA、Chip等。

21世纪初的微电子技术仍将以硅基CMOS电路为主流工艺，但将突破目前所谓的物理"限制"，继续快速发展；集成电路将逐步发展成为集成系统；微电子技术将与其他技术结合形成一系列新的增长点，例如微机电系统（MEMS）、

DNA芯片等。具体点说，SoC设计技术、超微细光刻技术、虚拟工厂技术、铜互连及低K互连绝缘介质、高K栅绝缘介质和栅工程技术、绝缘硅（Silicon On Insulator，SOI）技术等将在近几年内得到快速发展。因此，可以说，21世纪将是我国微电子产业的黄金时代。

微电子技术发展的目标是不断提高集成系统的性能及性能价格比，因此便要求提高芯片的集成度，这是不断缩小半导体器件特征尺寸的动力源泉。以MOS技术为例，沟道长度缩小可以提高集成电路的速度；同时缩小沟道长度和宽度还可减小元器件尺寸，提高集成度，从而在芯片上可以集成更多数目的晶体管，同时将结构更加复杂、性能更加完善的电子系统集成在一个芯片上；此外，随着集成度的提高，系统的速度和可靠性也大大提高，价格大幅度下降。

目前，0.18μm CMOS工艺技术已成为微电子产业的主流技术，0.035μm乃至0.020μm的器件已在实验室中制备成功，研究工作已进入亚0.1μm技术阶段，相应的栅氧化层厚度只有2.0～1.0nm。预计到2010年，特征尺寸为0.05～0.07μm的64GDRAM产品将投入批量生产。21世纪上半叶，微电子生产技术仍将以尺寸不断缩小的硅基CMOS工艺技术为主流。尽管微电子学在化合物和其他新材料方面的研究取得了很大进展，但还不具备替代硅基工艺的条件。根据科学技术的发展规律，一种新技术从诞生到成为主流技术一般需要20～30年的时间，硅集成电路技术自1947年发明晶体管到1958年发明集成电路，再到20世纪60年代末发展成为大产业也经历了20多年的时间。另外，全世界数以万亿美元计的设备和技术投入，已使硅基工艺形成非常强大的产业能力；同时，长期的科研投入已使人们对硅及其衍生物各种属性的了解达到十分深入、十分透彻的地步，成为自然界100多种元素之最，这是非常宝贵的知识积累。产业能力和知识积累决定了硅基工艺起码将在50年内仍起重要作用，人们不会轻易放弃。

当前，我国微电子技术产业与国际水平相比还属于初级阶段，无论技术水平、产品水平还是综合实力都无法与发达国家同行的实力相抗衡。回顾20世纪后50年，展望21世纪前50年，这百年的微电子科学技术发展历程，我们深切地感受到，世纪之交的微电子技术对我们既是一个重大的机遇，也是一个严峻的挑战；要抓住这个机遇，去勇敢地迎接这个挑战，使我国微电子技术实现腾飞，在新一代微电子技术中拥有自己的知识产权，促进我国微电子产业的发展。

本书由武汉通信指挥学院刘波主持翻译，总参第63研究所曲新波、解放军理工大学工程兵工程学院郭飞参与翻译，最后由武汉东风本田汽车有限公司肖文审校。同时感谢在本书翻译过程中给予过指导的武汉通信指挥学院的王祖平教授、姜文春教授、水海鹰教授、胡喜春副教授、陈卫东副教授、黎发新副教授、陈福中副教授，以及华为3Com公司的赵钱兵工程师、南京海脉科技有限公司的吴学智总经理、杭州世导科技有限公司的钟俊平工程师、华中科技大学的郎为民

博士、邓鹏、吴耀文、焦寨军、通信指挥学院的王汉杰、姚双庆、刘中治、周丹、张其增，感谢为本书作了大量文档和表格工作的肖文、沈斌、徐昶、丁飞、李国栋、莫春声。

需要说明的是，本书是译者在尽量忠实于原书的基础上翻译的，书中所述并不代表译者及其所在单位的观点。另外，为尽量保持原书特色，书中部分图形和文字符号并未按国家标准做修改，这点请读者注意。

由于时间仓促，加上译者水平所限，书中难免存在不妥之处，恳请广大读者批评指正。

译者

原书前言

微电子学在电子产业和电子系统领域中扮演着重要角色。在快速发展的重要电子应用领域中，小型化、快速化以及功耗越来越低、成本越来越低是永远不变的主旋律，因此，消费、工业以及军事领域中的电子产品对微电子技术的需求也在不断增长。随着制造业的不断发展，各种元器件的大批量生产技术越来越成熟，这样元器件的成本将不断下降。反过来，随着能耗的不断下降，各个产业也将朝着产品小型化和批量生产的方向发展。

微电子元器件尺寸越来越小、速度越来越快、容量越来越大的特点首先是被 Gordon E. Moore（著名的 Intel 前主席）在 20 世纪 60 年代发现的，他指出半导体晶体管的特征尺寸每年会减小 10%。而事实上，这种减小的速度比想象中要快得多。举例来说，动态随机存取存储器（Dynamic Random Access Memory, DRAM）集成电路的容量每三年大约翻两番。随着微电子元器件中晶体管集成度的不断提高，在实际应用中出现了“几乎不要计算功率”的现象。

20 世纪 80 年代爆发的所谓“信息时代”数字革命首先与桌面计算机联系在了一起，这场革命同时还带来了数据记录系统的真正发展。虽然那时向真正的数字系统过渡还有很远的距离，但是对以后数字系统的发展已经产生了深远的影响。数字系统中最重要的一点就是数字信息的通用性，即任何形式的符号可以通过量化处理转换成一条数字信息流，这样数据才能与其他符号串联在一起进行传输。

计算机可以合理地处理各种数据，还可以根据上下文来组织和存取信息。如今，计算机正在快速地改变着我们的世界——从办公室到家庭——从传统的独立式主机到嵌入式计算设备，几乎每一个设备或器件中都包含一个或多个微处理器。

微电子的市场需求已经由巨大的军用需求演变为巨大的消费市场需求。随之而来的是，电子产品的设计也开始向着迎合消费者需求的方向发展，例如低功耗、低成本以及大量市场应用产品的出现。而军用需求的发展则逐渐处于劣势，这是因为军用产品要求特定的可靠性、特定的外形、特殊的用途；这样一来，军用产品的成本就变得很高。因此，微电子元器件的性能无论是从产品的技术角度来看，还是从电子产品的实用性来看都非常重要，微电子元器件的发展目标就是使用户端产品以更高的效率完成比以前更为复杂的任务。

本书重点阐述了专用微电子技术中的技术要点，并探讨了它们如何推动下一

代微电子技术向前发展。本书中的各章主要围绕微电子技术的以下三个主题进行了详细阐述：材料、元器件和应用。

本书的主要目的是从微电子技术发展以及微电子技术对用户端产品的影响两个角度来为读者提供一个全面的阐述。

Jerry C. Whitaker

目　　录

译者序
原书前言

第1章　半导体材料 …… 1
1.1　引言 …… 1
1.2　晶体结构 …… 2
1.3　能带及相关半导体参数 …… 7
1.4　载流子的转移 …… 15
1.5　晶体缺陷 …… 20
1.6　小结 …… 24
参考文献 …… 25
第2章　热效应特性 …… 28
2.1　引言 …… 28
2.2　热效应基本原理 …… 28
2.3　其他材料特性 …… 31
2.4　工程数据 …… 32
参考文献 …… 34
第3章　半导体 …… 35
3.1　引言 …… 35
3.2　二极管 …… 36
参考文献 …… 38
第4章　MOS场效应晶体管 …… 40
4.1　引言 …… 40
4.2　伏安特性 …… 42
4.3　重要器件参数 …… 44
4.4　元器件微型化的局限性 …… 52
参考文献 …… 56
第5章　集成电路 …… 58
5.1　引言 …… 58
5.2　高速电路设计技巧 …… 58

参考文献 …… 67
第6章　集成电路设计 …… 69
6.1　引言 …… 69
6.2　IC设计过程概述 …… 69
6.3　IC设计总则 …… 70
6.4　小规模和中规模集成电路的设计 …… 76
6.5　LSI和VLSI电路设计 …… 83
6.6　MOS电路中不断增加的封装密度和不断减小的功耗 …… 84
6.7　门阵列 …… 86
6.8　标准单元 …… 87
6.9　可编程逻辑器件 …… 87
6.10　不断减小的传输延时 …… 91
6.11　输出缓冲器 …… 94
参考文献 …… 96
第7章　数字逻辑系列 …… 98
7.1　引言 …… 98
7.2　晶体管-晶体管逻辑电路 …… 99
7.3　CMOS逻辑电路 …… 104
7.4　发射极耦合逻辑电路 …… 109
7.5　可编程逻辑电路 …… 113
参考文献 …… 117
第8章　存储器 …… 119
8.1　引言 …… 119
8.2　存储器构造 …… 120
8.3　存储器类型 …… 124
8.4　接口存储器 …… 136
8.5　检错与纠错 …… 141
参考文献 …… 142
第9章　微处理器 …… 144
9.1　引言 …… 144
9.2　体系结构的基本要素 …… 144
参考文献 …… 154
第10章　D/A和A/D转换器 …… 156
10.1　引言 …… 156
10.2　D/A和A/D转换电路 …… 156

参考文献 …… 164
第 11 章 专用集成电路 …… 166
11.1 引言 …… 166
11.2 全定制 ASIC …… 167
11.3 半定制 ASIC …… 168
参考文献 …… 183
第 12 章 数字滤波器 …… 185
12.1 引言 …… 185
12.2 FIR 滤波器 …… 185
12.3 IIR 滤波器 …… 192
12.4 有限字长效应 …… 202
参考文献 …… 211
第 13 章 多片组件设计技术 …… 214
13.1 引言 …… 214
13.2 多片组件设计技术的定义 …… 214
13.3 设计、修补和测试 …… 222
13.4 MCM 的应用 …… 225
13.5 MCM 的设计要点 …… 226
参考文献 …… 228
第 14 章 集成电路的测试 …… 230
14.1 引言 …… 230
14.2 缺陷类型 …… 230
14.3 测试的概念 …… 232
14.4 测试中的权衡 …… 238
参考文献 …… 239
第 15 章 半导体故障模式 …… 241
15.1 分立半导体故障模式 …… 241
15.2 集成电路故障模式 …… 246
15.3 混合微电路及故障 …… 247
15.4 存储 IC 故障模式 …… 248
15.5 IC 封装及故障 …… 251
15.6 除铅 …… 253
15.7 筛选测试及再筛选测试 …… 254
15.8 静电放电效应 …… 257
参考文献 …… 260

第 16 章 基本计算机体系结构 …… 262
16.1 引言 …… 262
16.2 计算机体系结构的定义 …… 262
16.3 单处理器系统 …… 263
16.4 多处理器系统 …… 267
16.5 存储器分级体系 …… 269
16.6 实现过程中的注意事项 …… 270
参考文献 …… 278
第 17 章 软件设计与开发 …… 279
17.1 引言 …… 279
17.2 软件的概念 …… 279
17.3 软件工程的实质 …… 284
17.4 一种新的设计方式 …… 289
17.5 Apollo 板上飞行软件成果：值得学习的课题 …… 291
17.6 "事先预防"式开发 …… 294
17.7 "事先预防"理论 …… 297
17.8 设计及开发过程 …… 298
17.9 选择正确的模式并实现自动操作 …… 301
参考文献 …… 304
第 18 章 神经网络与模糊系统 …… 307
18.1 神经网络与模糊系统 …… 307
18.2 神经单元 …… 307
18.3 前馈神经网络 …… 310
18.4 神经网络的学习算法 …… 311
18.5 特殊前馈网络 …… 317
18.6 循环神经网络 …… 323
18.7 模糊系统 …… 325
18.8 设计实例 …… 328
18.9 遗传规则方法 …… 329
参考文献 …… 332
第 19 章 机器视觉 …… 334
19.1 引言 …… 334
19.2 成像过程 …… 336
19.3 分割 …… 341
19.4 特征提取和匹配 …… 346

19.5 三维目标识别 …… 351
19.6 动态视觉 …… 355
19.7 应用 …… 359
参考文献 …… 363
第20章 语音增强技术概述 …… 365
20.1 引言 …… 365
20.2 信号子空间法 …… 366
20.3 短期频谱估计 …… 369
20.4 多状态语音模型 …… 371
20.5 二阶统计估计 …… 372
20.6 结束语 …… 376
参考文献 …… 376
第21章 Ad Hoc 网络 …… 378
21.1 引言 …… 378
21.2 路由算法 …… 378
21.3 媒体接入协议 …… 384
21.4 Ad Hoc 网络的 TCP …… 395
21.5 Ad Hoc 网络的容量 …… 398
参考文献 …… 402
第22章 网络通信 …… 407
22.1 网络分析与设计的一般原则 …… 407
22.2 个人远程连接 …… 411
22.3 局域网 …… 413
22.4 互联组网 …… 425
22.5 广域网 …… 432
参考文献 …… 436
第23章 印刷技术及系统 …… 439
23.1 引言 …… 439
23.2 印刷技术 …… 442
23.3 非击打式印刷技术 …… 442
23.4 热印刷技术 …… 450
23.5 电子照相印刷技术 …… 452
23.6 磁成像和离子成像技术 …… 455
23.7 系统要点 …… 455
参考文献 …… 459

第 24 章　数据存储系统 …… 461
24.1　引言 …… 461
24.2　独立磁盘冗余阵列系统 …… 461
参考文献 …… 469
第 25 章　光学存储系统 …… 470
25.1　引言 …… 470
25.2　光度头 …… 470
25.3　WORM 技术 …… 478
25.4　磁-光技术 …… 478
25.5　可记录压缩盘 …… 480
25.6　光学磁盘系统 …… 483
参考文献 …… 488
第 26 章　纠错技术 …… 490
26.1　背景知识 …… 490
26.2　引言 …… 491
26.3　纠错码的发展 …… 492
26.4　纠错码系列 …… 493
26.5　线性块状码 …… 494
26.6　卷积码 …… 497
26.7　格状编码调制 …… 499
26.8　应用 …… 500
参考文献 …… 502
缩略语 …… 504

第1章 半导体材料

StuartK. Tewksbury

1.1 引言

半导体是指导电能力介于导体和绝缘体之间的材料。与电阻材料中的颗粒状结构不同，半导体材料具有晶体结构，其导电性能是通过半导体原子中的复杂量子力学特性来确定的，这些原子在半导体中是呈周期性排列的。对于很多元素来说，晶体结构使原子的受束缚电子和自由电子（例如，不受束缚于原子的电子）能级之间产生了一个能带，该能带直接影响着晶体中与原子相关的电子成为载流子（即自由电子）的机制。半导体的导电性能与载流子的密度成正比，自由载流子密度的变化范围很广；我们可以将本征半导体中的很小一部分原子（大约百万分之一）替换成其他不同类型的原子（即掺杂原子），这样就可以改变自由载流子的密度。多数载流子的密度与掺入的杂质密度直接相关。通过在晶体中选择性地掺入其他元素，晶体就可以实现不同的导电性能。另外，载流子密度（自由电子的密度）由掺入的施主杂质决定，而与自由电子相对应空穴的密度则由掺入的受主杂质决定（在这类半导体中空穴数是自由电子数的两倍）。因此，可以通过掺入杂质的类型来区分半导体的类型（N型半导体：自由电子密度远远大于空穴密度；P型半导体：空穴密度远远大于自由电子密度）。

如果施加合适的电场，半导体中的中间微小区域可以处于一种所有载流子（电子和空穴）都被电场驱逐的状态，而电场则由裸露的掺杂离子维持。这样，在导电状态（可变的电阻率）和非导电状态（导电性随着载流子的消失而消失）之间就可以形成一个电子开关。

对具有精确控制电阻率（由电子或空穴控制）的局部区域进行合并，以及实现对载流子（电子或空穴）的控制，这两点构成了半导体技术的基本原理，同时也是当前电子技术领域的基础。这个基础是非常牢固的，因为原子元素种类很多（包括不同原子的混合），这样我们就可以根据不同的需求而采用不同的半导体材料。硅半导体材料在电子技术领域（例如：超大规模集成（VLSI）数字电子领域）中占据了绝对优势，这与光电子学中半导体材料的多样性形成了鲜明对比。在光电子学中，我们可以通过调整“能带间隙”宽度来获得不同波长的光波，这一点直接导致了大量基于不同技术的光电子器件的诞生。非硅半导体

技术在电子元器件中同样占据重要地位，尤其是在那些采用了类似于光电子技术中的原子元素来实现高速半导体的高带宽电路中。因此，随着光电子技术的不断发展，非硅半导体技术的重要性将会不断地提高，这是因为一个很简单的原因：硅不适合制造高效率的光源。

1.2 晶体结构

基本半导体材料族

大多数半导体材料都是晶体结构，晶体结构是由原子的共价键形成的。在共价键的作用下，每个原子的价电子层都被8个电子填充，其中4个电子来自于相邻的4个原子，每个原子提供1个共价电子。半导体材料中包含了由单一元素构成的半导体，在这些半导体中每个元素原子的价电子层拥有4个电子（剩下的4个由相邻原子的共价键来补充），元素来自于元素周期表中的第Ⅳ组。而其他半导体材料都是由两种元素组成的，其中一种来自于元素周期表中的第N组（$N<4$），另一种来自于元素周期表中的第M组（$M>4$）；这里，$N+M=8$，价电子层由8个电子填充。

半导体材料的主要种类归纳如下。

1. 本征半导体（Ⅳ~Ⅳ）

本征半导体是指晶体结构由元素周期表第Ⅳ组中的单一元素构成的半导体，这些元素包括：锗（Ge）、硅（Si）、碳（C）和锡（Sn）。其中，硅在电子半导体器件中应用最广泛，而且几乎是地球上最常见的元素。表1-1归纳了半导体中常用的部分元素在自然界中的含量，其中包括了非本征（化合物）半导体。

表1-1 半导体中常用元素含量（地球上现有的部分元素）

元素	含量	元素	含量
Si	0.28	Ge	5×10^{-6}
Ga	1.5×10^{-5}	Cd	2×10^{-7}
As	1.8×10^{-6}	In	1×10^{-7}

图1-1a描述了共价键形成的原理（共用两个原子的价电子层电子）。通过共价键，晶体中的第Ⅳ组原子与相邻的原子紧紧吸引在一起，并将价电子层填满，形成一个稳定的分子网络。

半导体晶体除了可以由元素周期表中第Ⅳ组的单一元素构成外，也可以由第Ⅳ组中的两种或多种不同元素构成。例如，由硅和碳组成的硅碳化合物（SiC）已经用于高温环境中了。另外，用来实现能带间隙工程的Si_xGe_{1-x}化合物正处在研究之中。其中，x表示该化合物中Si元素所占的比例（$0<x<1$），而剩下$1-x$的部分由Ge元素代替。这种由周期表中同一组的两种元素组合来代替单一元素的

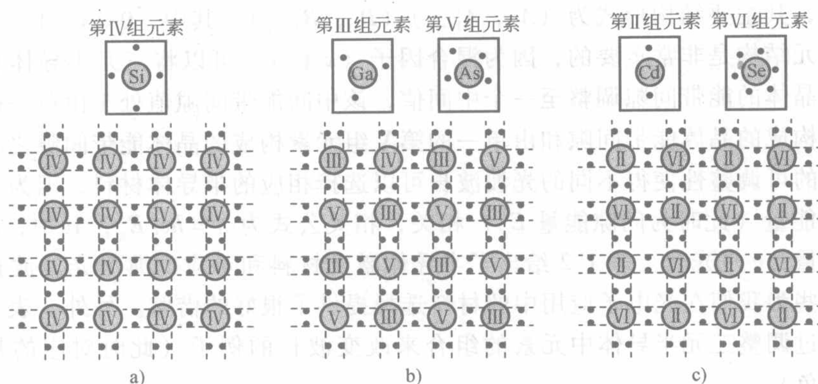


图 1-1　半导体晶体结构中原子的紧密排列示意图
a）硅半导体
b）第Ⅲ组与第Ⅴ组元素构成的化合物半导体 GaAs
c）第Ⅱ组与第Ⅵ组元素构成的化合物半导体 CdS

做法在其他类型的半导体中也出现了，具体描述如下。

（1）Ⅲ-Ⅴ化合物半导体

Ⅲ-Ⅴ化合物半导体在光电子元器件的应用中占有重要地位（将来也会越来越重要）。另外，对于电子产品来说，Ⅲ-Ⅴ化合物半导体在运行速度上比硅半导体更具潜力，尤其是在无线通信领域。化合物半导体具有晶格结构，该晶格结构由周期表中的不同组元素原子构成。例如，Ⅲ-Ⅴ化合物半导体就是由第Ⅲ组的元素 A 和第Ⅴ组的元素 B 构成。值得注意的是，每一个第Ⅲ组中的元素原子只能与 4 个第Ⅴ组中的元素原子组合，而每一个第Ⅴ组中的元素原子也只能与 4 个第Ⅲ组中的元素原子组合，具体组合如图 1-1b 所示。其中，原子之间的结合是通过共用电子来形成的，这样一来，每个原子都有一个被填满的价电子层（8 个电子）。通过共用电子形成的这种结合称为“共价键”，共价键是很稳定的，不会因为产生共用电子的原子由第Ⅲ组变成第Ⅴ组，或者其中甚至还包含离子键（相比由纯共价键构成的半导体而言）而改变。其中，具有代表性的Ⅲ-Ⅴ化合物半导体包括：GaP、GaAs、GaSb、InP、InAs 和 InSb。

GaAs 可能是最常见的Ⅲ-Ⅴ化合物半导体，它主要用于制造高速电子元器件和光电子元器件。光电子元器件主要利用三元和四元Ⅲ-Ⅴ化合物半导体来实现各种不同的光波波长，并形成各种新奇复杂的元器件结构。三元半导体的结构形式主要包括 $(A_x, A'_{1-x})B$（主要由第Ⅲ组中的两个不同元素原子构成，这两个原子用来填充晶格中第Ⅲ组元素原子的位置）和 $A(B_x, B'_{1-x})$（主要由第Ⅴ组中的两个不同元素原子构成，这两个原子用来填充晶格中第Ⅴ组元素原子的位置）两种。四元半导体的结构主要由两个第Ⅲ组元素原子和两个第Ⅴ组元素原

子组成，其主要结构形式为（A_x，A'_{1-x}）（B_y，B'_{1-y}）。其中，$0 \leqslant x \leqslant 1$。这种三元和四元结构是非常重要的，因为混合因子（$x$ 和 y）可以将三元半导体和四元半导体晶体的能带间隙调整至一个中间值，该中间能带间隙值处于由单一的第Ⅲ组元素构成的晶体能带间隙和由单一的第Ⅴ组元素构成的晶体能带间隙之间。光波波长的可调整性使得不同的光波波长可以选择相应的半导体材料，因为光波波长 λ 与能量（此时为间隙能量 E_g）相关，相关公式为 $\lambda = hc/E_g$；其中，h 是普朗克常量，c 是光速。表 1-2 给出了半导体激光材料和对应的典型光波波长的例子，这些为我们在光电子应用中的材料选择提供了很好的借鉴。另外，表 1-3 给出了通过调整三元半导体中元素的组合来改变波长的例子（此时对应的是可见光的颜色）。

表 1-2 半导体光源和其典型波长

激光使用材料	波长/nm	激光使用材料	波长/nm
ZnS	454	InGaAsSb/GaSb	2200
AlGaInP/GaAs	580	AlGaSb/InAsSb/GaSb	3900
AlGaAs/GaAs	680	PbSnTe/PbTe	6000
GaInAsP/InP	1580		

表 1-3 $GaAs_xP_{1-x}$ 半导体中调整波长的变量 x 示例

三元化合物	颜色	三元化合物	颜色
$GaAs_{0.14}P_{0.86}$	黄	$GaAs_{0.6}P_{0.4}$	红
$GaAs_{0.35}P_{0.65}$	橙		

与由单一元素构成的半导体（晶格点阵中每个原子的定位是独立的）相比，Ⅲ-Ⅴ化合物半导体要求在晶体生长过程中对化学计量（例如，两种不同原子的比例）具有很好的控制。举例来说，每个镓（Ga）原子必须紧挨着一个镓原子（而不是砷（As）原子），反之亦然。因此，通常大型Ⅲ-Ⅴ高质量晶体的生长会比单一元素构成的晶体（例如，硅（Si））生长要困难一些。

（2）Ⅱ-Ⅳ化合物半导体

Ⅱ-Ⅳ化合物半导体晶体是由元素周期表中第Ⅱ组的一个原子与第Ⅳ组的一个原子进行组合构成的，其中每一种类型的原子必须和 4 个相邻的另一种类型的原子结合，如图 1-1c 所示。从第Ⅳ组到第Ⅱ组原子逐渐增加的电荷总量来看，该类型半导体中的共价键比Ⅲ-Ⅴ化合物半导体中的共价键更趋向于离子键。与Ⅲ-Ⅴ化合物半导体结构类似的是，Ⅱ-Ⅳ化合物半导体也可以产生三元和四元结构。尽管没有Ⅲ-Ⅴ化合物半导体常见，但是Ⅱ-Ⅳ化合物半导体主要用来满足一些非常重要应用的需求。其中，具有代表性的Ⅱ-Ⅳ化合物半导体包括：ZnS、ZnSe 和 ZnTe（形成的闪锌矿晶格结构将在接下来讨论）、CdS 和 CdSe（形成闪锌矿或纤维锌矿晶格结构）以及 CdTe（形成纤维锌矿晶格结构）。

2. 三维晶格结构

前面的二维结构示意图已经详细描述了共价电子以及共价键形成的原理。然而，三维结构的原理要比二维结构复杂得多。不过，绝大多数的半导体晶体都具有一个通用的基本结构，具体描述如下。

最原始的晶体结构是由 8 个原子构成的立方体，如图 1-2a 所示。对这种结构进行扩展时，可以在立方体每个面的中心增加一个原子，这样就变成了一个以面为中心的立方体结构（每个晶格就包含 14 个原子），如图 1-2b 所示。晶格常数就是指立方体边的长度。

全晶格结构是指包含两个面心立方体结构（FCC）晶格的结构；其中，一个晶格穿插在另一个晶格之中（例如，一个晶格立方体的一个角插入另一个晶格立方体之中，同时两个晶格的各个面保持平行），如图 1-2c 所示。在基于这种 FCC 晶格结构的Ⅲ-Ⅴ化合物半导体中，其中一个 FCC 晶格由某一种类型的元素原子构成（例如第Ⅲ组类型），而另一个 FCC 晶格则由另外一种类型的元素原子构成（例如第Ⅴ组类型）。对于三元和四元半导体而言，来自同一组的元素原子位于同一个 FCC 晶格结构中。不同晶格的原子之间也同样存在共价键。例如，GaAs 晶体中的所有镓（Ga）原子都处于一个 FCC 晶格结构中，它会与处于另一个 FCC 晶格结构中的砷（As）原子形成共价键。因此，相邻原子之间的距离就比晶格常数要小。例如，硅原子的原子间距为 2.35Å㊀，而硅的晶格常数是 5.43Å。

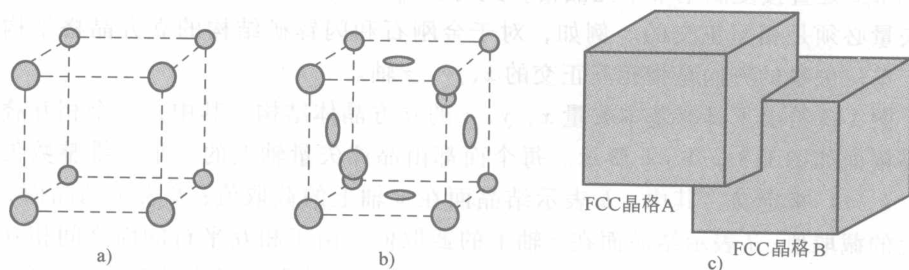


图 1-2　三维晶格结构

a）基本立方体结构　b）面心立方体（FCC）结构　c）两个相互穿插的面心立方体（FCC）结构

注：图中的虚线并不是指原子间的共价键，而仅仅是用来描述立方体外形的轮廓

在全晶格结构中，如果组成两个晶格的元素是来自周期表的不同元素组，那么这种晶体结构被称为“闪锌矿晶格”。在单一元素构成的半导体中（如硅半导体），FCC 晶格中都是硅原子，这种晶格结构称为“金刚石晶格”（碳原子通过

㊀ 1Å（埃）$=1\times10^{-10}$m。

结晶成金刚石晶格形成了真正的钻石晶体，而碳正是第Ⅳ组的元素)。正如前面讨论的Ⅲ-Ⅴ化合物半导体一样，在硅晶体中的硅原子之间形成的共价键同样扩展到了FCC晶格的子晶格中。

虽然常见的半导体材料大多数都拥有上面描述的基本金刚石或闪锌矿结构，但也有一些半导体晶体是六角紧密（HCP）晶格结构，例如CdS和CdSe。在CdS和CdSe晶体中，所有的镉（Cd）原子位于一个HCP晶格上，而所有的硫（S）原子或硒（Se）原子都位于另一个HCP晶格上。在金刚石和闪锌矿结构晶体内部，一个完整的晶格是由两个相互穿插的HCP晶格组成的。所有这些晶体结构统称为“纤维锌矿晶格”。Ⅳ-Ⅵ类型的化合物半导体（PbS、PbSe、PbTe和SnTe）可以很好解释窄能带间隙的原理，并且已经用于红外探测器中；而上面这些Ⅳ-Ⅵ化合物半导体的晶格就是简单的立方体晶格结构（也称为“NaCl”晶格）。

3. 晶体结晶方向和结晶面

晶体的结晶方向和结晶面对于半导体的电气特性和实际应用非常重要，因为不同的结晶面表现出的物理特性大不相同。例如，原子的表面密度（每平方厘米[⊖]的原子数）在不同的结晶面会大不一样。通常，我们采用一个规范的参数来定义结晶面和正常的结晶方向，即所谓的“米勒指数”。

通常，晶格定义了一组单位矢量（例如a、b和c）；沿着单元晶格矢量的方向复制整数倍的单元晶格就可以构建一个完整的晶体；也就是说，在位置$n_a a + n_b b + n_c c$处直接复制基本单元晶格，此时，n_a、n_b和n_c都是整数。一般来说，单位矢量必须是相互正交的。例如，对于金刚石和闪锌矿结构的立方晶格结构来说，单位矢量的方向是指相互正交的x，y，z轴。

图1-3给出了具有基本矢量x、y、z的立方晶体结构。其中，三个相互叠加的晶格面如图1-3a、b、c所示。每个面都由晶体矢量轴上的一组三维整数矢量（h，k，l）来定义。其中，h表示结晶面在x轴上的截取值；k表示结晶面在y轴上的截取值；l表示结晶面在z轴上的截取值。由于相互平行的面之间相互等价，因此截取的整数值取具有相同比例关系的三维整数矢量中的最小值。例如，(100)、(010)和(001)对应的就是立方体的三个面，而(111)对应的则是在x、y、z轴上的截取值分别为(1，1，1)的面。如果截取值为负数，那么相应的米勒指数则表示成一个整数上面加一个横杠，例如($\bar{1}00$)，这个面类似于(100)面，但是与x轴相交于数值-1处。

附加指数用来表示对称面的矢量以及方向，例如{100}表示等价面(100)($\bar{1}00$)(010)($0\bar{1}0$)(001)和($00\bar{1}$)。(hkl)面的正常方向应该是[hkl]。不

⊖ 1平方厘米（$1cm^2$）$=1\times10^{-4}m^2$。

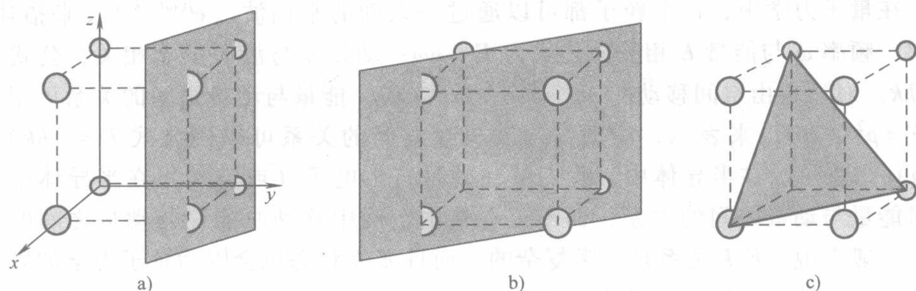


图 1-3　碳晶体的立方体型晶体结晶面示例

a)（010）面　b)（110）面　c)（111）面

同的结晶面在元器件的生产过程中表现出来的特性也是不同的，并且会严重影响到元器件的电气特性。一个不同之处就是，晶体晶格表面的不同重构会将能量最小化；另一个不同之处是，不同的结晶面会产生不同的原子表面密度。例如，硅晶体中结晶面（100），（110）和（111）对应的原子表面密度（每平方厘米上的原子数）分别是 6.78×10^{14}、9.59×10^{14}和 7.83×10^{14}。

1.3　能带及相关半导体参数

半导体晶体中的原子具有周期性的排列，这种排列使得晶体内部的势能具有周期性空间变化特性。由于这种潜在的能量（即势能）会随着原子间距的变化而产生很大的变化，因此，在研究半导体能级以及相关特性时，必须以量子力学原理为基础。不同的半导体晶体（不同的原子种类以及不同的原子间距），它们的特性是不一样的。但是，这种势能变化的周期性带来的结果是所有的半导体晶体都具有一些“一般特性”；除了这些一般特性外，不同的半导体材料还会表现出一些与“一般特性”相关的特性。关于上面这些量子力学结果的详细研究不在本小节探讨的范围之内，因此，我们只能假设这种“一般特性”。

对于半导体材料的研究来说，一个主要的研究内容就是能量-动量函数，该函数定义了电荷载流子的状态。除了我们熟悉的自由电子外，半导体材料中还包含了与自由电子具有类似特性的空穴载流子（带正电的粒子）。这样，在半导体中就出现了两个重要的能级：一个是以晶体内自由移动而不受原子束缚的自由电子为基础的能级（导带），另一个就是以空穴为基础的能级（价带），在这两个能级之间有一个能量禁区（不存在任何自由载流子）。导带能级与价带能级之间最小的间距称为“能带间隙”或“能隙”，能带和能隙都是半导体材料的重要基本特性。

1. 导带和价带

在量子力学中，每个粒子都可以通过一系列的平面波（$e^{j(\omega t-kx)}$）来描述；其中，频率 ω 与能量 E 相关，公式为 $E=\hbar\omega$；动量 p 与波矢量 $\boldsymbol{k}$ 相关，公式为 $\boldsymbol{p}=\hbar\boldsymbol{k}$。对于自由空间移动的 m 个典型粒子来说，能量与动量之间的关系可用公式 $E=p^2/(2m)$ 来表示，而动量与波矢量之间的关系可以用公式 $E=(\hbar\boldsymbol{k})^2/(2m)$ 来表示。在半导体中，我们关注的是自由电子（或空穴）在半导体中移动时能量与动量之间的关系，而不是在自由空间中移动时能量与动量之间的关系。一般来说，E-$\boldsymbol{k}$ 关系是非常复杂的，而且 E-$\boldsymbol{k}$ 状态也会因为量子力学的影响而存在多种情况。例如，晶体中原子位置的周期性导致波矢量 $\boldsymbol{k}$ 也是周期性的，同时要求 $\boldsymbol{k}$ 的值处于一定范围之内的（根据 E-$\boldsymbol{k}$ 关系）。

图 1-4 给出了一个能量-动量面（例如，E-$\boldsymbol{k}$ 面）的导带和价带的例子（假设情况）。导带的 E-$\boldsymbol{k}$ 关系图中有一个最小的能量值，在平衡条件下，电子将更趋向处于该最小能量状态。导带中也存在高于这个最小能量值（E_c）的电子能级，它们各对应了一个动量值。价带的 E-$\boldsymbol{k}$ 关系对应的是空穴的能量-动量关系。这样，在关系图中越向下，能量值越高，而且在图 1-4 中最小价带能级 E_v 是最大的能量值。当共价键中的一个价电子获得足够的能量（受激发）而挣脱原子的束缚成为自由电子的同时，在共价键中会留下了一个空位，即“空穴”。因此，能隙能量 $E_g=E_c-E_v$ 就表示了产生一个电子-空穴对时所需的最小能量（更高的能量会产生能量值比 E_c 更大的自由电子，但是这些电子会逐渐失去多余的能量回到最小值的能量状态）。

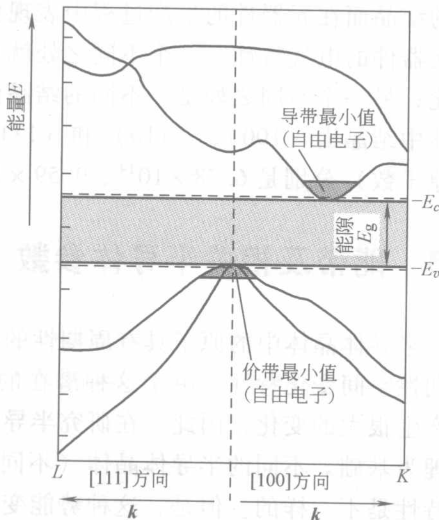


图 1-4 导带和价带一般结构

能带和能隙的具体特性与半导体晶体中的量子力学结构直接相关。晶体结构（即使是硅半导体晶体）的变化会直接导致能带发生变化。由于半导体的热扩展系数是非零的，因此温度的改变会直接导致原子间距发生变化，从而导致能隙与温度产生直接关系。另外，外部压力的变化也会导致原子间距发生变化；尽管这些变化都是很微小的，但仍然可以从能隙值反映出来。表 1-4 给出了在室温条件下几种常见半导体的能隙值 E_g，同时该表中还有 E_g 随温度（T）变化和 E_g 在室温下随压力（P）变化的关系。

表 1-4　能带随温度和压力的变化

半导体	E_g(300K)	dE_g/dT/(meV/K)	dE_g/dP/(meV/kbar)
Si	1.110	-0.28	-1.41
Ge	0.664	-0.37	5.1
GaP	2.272	-0.37	10.5
GaAs	1.411	-0.39	11.3
GaSb	0.70	-0.37	14.5
InP	1.34	-0.29	9.1
InAs	0.356	-0.34	10.0
InSb	0.180	-0.28	15.7
ZnSe	2.713	-0.45	0.7
ZnTe	2.26	-0.52	8.3
CdS	2.485	-0.41	4.5
CdSe	1.751	0.36	5
CdTe	1.43	-0.54	8

来源：Böer，K. W. 1990. *Survey of Semiconductor Physics*，*Vol. 1*：*Electrons and Other Particles in Bulk Semiconductors.* Van Nostrand，New York.

随温度变化的关系虽然不是很明显，但是对载流子的密度会产生很大的影响。一个 E_g 随温度变化的启发模型为 $E_g(T) = E_g(0K) - \alpha T^2/(T+\beta)$。在这个公式中，各个参数的值如表 1-5 所示。对于 GaAs 的能隙，在 0 ~ 1000K 之间时，由上述公式得出的值可以精确到 2×10^{-3}eV。

表 1-5　常见半导体中与温度相关的参数

	E_g(0K)/eV	$\alpha(\times10^{-4})$	β	E_g(300K)/eV
GaAs	1.519	5.405	204	1.42
Si	1.170	4.73	636	1.12
Ge	0.7437	4.774	235	0.66

来源：Sze，S. M. 1981. *Physics of Semiconductor Devices*，2nd ed. Wiley-Interscience，New York.

2. 直接能隙和间接能隙半导体

图 1-5 给出了 Ge、Si 和 GaAs 晶体的能带示意图。在图 1-5b 中，硅晶体的价带能量在动量值为 k 时达到最小，与导带取最小值时的 k 值不同。这就是一个“间接能隙”，“间接能隙”会产生一个电子-空穴对，同时也需要一个能量 E_g，而且在动量上也会产生相应的变化（例如 k）。在电子-空穴对的直接复合过程中，同样会在动量上产生相应的变化。这种对动量变化的需求（遵循能量和动量守恒定律）导致了在载流子对的直接复合过程中需要声子的参与，从而导致了“声子”的产生。将硅晶体直接作为光电子原料，这是不切实际的。但是，将硅和其他的“直接能隙”半导体结合在一起作为光学探测器材料时，这个粒子产生过程就变得非常容易了（同时产生 1 个电子、1 个空穴和 1 个声子）。

对于图 1-5c 中的 GaAs 晶体来说，导带最小值和价带最小值都处于同一个动

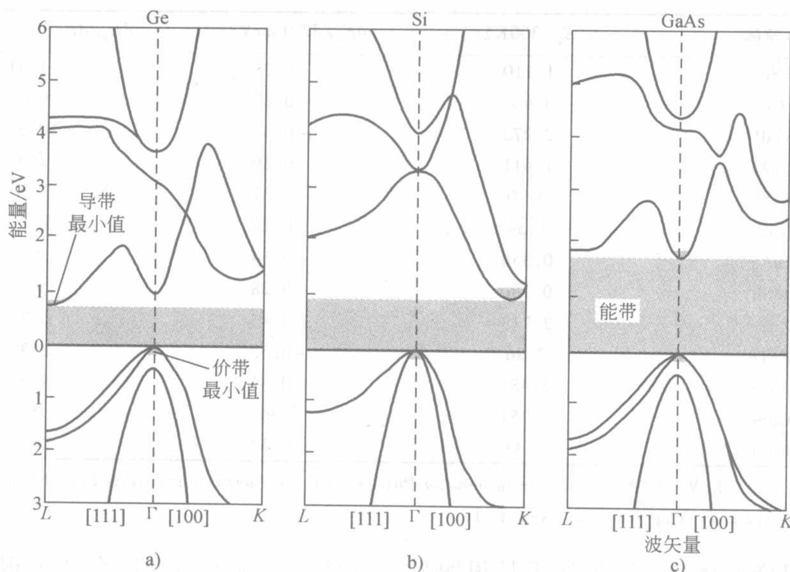


图 1-5　Ge、Si 和 GaAs 晶体中的导带和价带（来源：Sze，S. M. 1981. *Physics of Semiconductor Devices*, 2nd ed. Wiley Interscience, New York）

a) Ge　b) Si　c) GaAs

量值 k 处，这就是一个“直接能隙”。由于在直接复合过程中不发生动量变化，因此这样的复合过程非常容易，同时还会产生一个带有原始电子和空穴能量的声子（例如，一个声子的能量等于能隙值）。因此，直接能隙半导体就成为了有效的光学原料（各种直接能隙半导体的不同能隙值 E_g 可以用来实现各种不同范围的波长）。波长 λ 与能隙值 E_g 之间的关系可以用下面的公式来描述：$\lambda = hc/E_g$。

图 1-5c 还给出了间接能隙下导带能量的次最小值，该值比相应直接能隙下导带能量的最小值要大。更高的导带最小值可以由电子产生（电子处于更高能量的平衡状态），但是电子总数会减少，因为部分电子获得了足够的能量，冲破了外层的束缚。

3. 载流子的有效质量

对于一个能量非常接近导带最小值能量的电子，其能量随波矢量变化的能量-波矢量（E-$\boldsymbol{k}$）关系可以用方程式 $E(\boldsymbol{k}) = E_0 + a_2(\boldsymbol{k} - \boldsymbol{k}^*)^2 + a_4(\boldsymbol{k}-\boldsymbol{k}^*)^4 + \cdots$ 来描述。其中，$E_0 = E_c$ 是自由电子静止时的基态能量，$\boldsymbol{k}^*$ 是指产生导带最小值时的波矢量。在 $E(\boldsymbol{k})$ 的扩展项中，只包含与 $\boldsymbol{k}^*$ 有关的 $\boldsymbol{k}$-$\boldsymbol{k}^*$ 幂次项，因为 E-$\boldsymbol{k}$ 关系图是以 $\boldsymbol{k} = \boldsymbol{k}^*$ 为中心对称的。这样，E 的近似值就可以用来描述 E_c 上的微小增量。当增量足够小时，$\boldsymbol{k}$-$\boldsymbol{k}^*$ 的二次以上幂项就可以忽略不计，从而 $E(\boldsymbol{k}) \approx$

$E_0+a_2\boldsymbol{k}^2$，此时我们假设 $\boldsymbol{k}^*=0$。如果自由电子是在自由空间移动而不是在晶体中移动，而且其能量为 E_0，那么能量-动量（E-$\boldsymbol{k}$）关系可以描述为 $E(\boldsymbol{k})=E_0+(\hbar k)^2/(2m_0)$，这里 m_0 是指一个电子的质量。通过比较以上结果，我们可以将与导带抛物线最小值有关的曲率系数 a_2 和有效质量 m_e^* 联系起来，如下式所示：$a_2=(\hbar)^2/(2m_e^*)$ 或者

$$\frac{1}{m_e^*}=\frac{2}{\hbar^2}\frac{\partial^2 E_c(k)}{\partial k^2}$$

类似地，空穴的有效质量 m_h^* 也可以由价带抛物线最小值的抛物线系数来定义，如下式所示

$$\frac{1}{m_h^*}=\frac{2}{\hbar^2}\frac{\partial^2 E_v(k)}{\partial k^2}$$

由于能带与温度和压力相关，因此尽管在元器件的运行过程中室温和压力都是正常值，但它们仍然可以作为有效质量的属性。

以上讨论是基于一个假设的简单标准变量 k，而实际上，波矢量 $\boldsymbol{k}$ 是一个三维变量（k_1，k_2，k_3），其中，矢量方向由晶体的单位矢量定义。这样，对每个矢量 $\boldsymbol{k}$ 来说就包括 3 个独立的有效质量：m_1、m_2、m_3。每个标准的有效质量 m^* 都可以由这些不同方向的质量来定义，其相互关系与不同方向的质量有关。对于碳晶体（例如金刚石和闪锌矿结构）来说，矢量方向都是相互正交的，而且 $m^*=(m_1m_2m_3)^{1/3}$。如果该等式的 3 个变量中有两个值相等（例如 $m_1=m_2$），那么等式就变成为由两个变量构成，如同硅晶体和其他晶体中纵横结构的有效质量矢量（对应 m_l 和 m_t）。此时，$m^*=[(m_t)^2m_l]^{1/3}$。如果 3 个矢量 m_1、m_2、m_3 的值都相等，那么等式就可以直接使用简单的 m^* 值了。

在图 1-5 的价带结构中我们还可以注意到，两个不同的 E-k 价带可以具有相同的最小值。这是由于它们的抛物线系数不同，两个价带对应了不同的质量矢量，其中一个对应较重空穴的 m_h，另外一个对应较轻空穴的 m_l，此时有效质量为 $m^*=(m_h^{3/2}+m_l^{3/2})^{2/3}$。这种较重和较轻的空穴在很多半导体中都存在，包括硅晶体。有效质量的值如表 1-8 所示。

4. 本征载流子密度

导带自由电子的密度与两个函数相关，其中一个是电子的状态密度函数 $D(E)$，另一个是自由电子的能量分布函数 $F(E,T)$。

能量分布函数（热平衡条件下）由费米-迪拉克（Fermi-Dirac）分布函数定义：

$$F(E)=\left[1+\exp\left(\frac{E-E_f}{k_BT}\right)\right]^{-1}$$

在实际情况中，该公式可以近似为经典的麦克斯韦-波耳兹曼（Maxwell-Boltzmann）分布函数。该分布函数是通用的函数，与半导体的材料无关。

而状态密度函数 $D(E)$ 与半导体的材料有关，通常近似为

$$\text{电子}:D_n(E)=M_c\frac{\sqrt{2}}{\pi^2}\frac{(E-E_c)^{1/2}}{\hbar^3}\frac{(m_e^*)^3}{2}$$

$$\text{空穴}:D_p(E)=M_v\frac{\sqrt{2}}{\pi^2}\frac{(E_v-E)^{1/2}}{\hbar^3}\frac{(m_h^*)^3}{2}$$

式中，M_c和 M_v分别为具有相同最小导带值和价带值的数量。

值得注意的是，自由电子对应的 $E\geqslant E_c$区域和空穴对应的 $E\leqslant E_v$区域与 E_c和 E_v之间的能隙密切相关。

导带中电子的密度 n 可以由下式计算：

$$n=\int_{E=E_c}^{\infty}F(E,T)D(E)\mathrm{d}E$$

对于典型的处于 E_c之下和 E_v之上的费米（Fermi）能级（高于 k_BT）来说，上面积分的结果为

$$n=N_c\mathrm{e}^{-(E_c-E_f)/k_BT};$$
$$p=N_v\mathrm{e}^{-(E_f-E_v)/k_BT}。$$

式中，n 和 p 分别为导带自由电子密度和价带空穴密度；N_c和 N_v是指随温度、有效质量和其他条件变化而变化的“有效状态密度”（变化慢于前面公式中的指数）。

表 1-6 给出了部分半导体的 N_c和 N_v值，其公式分别为 $N_c=2(2\pi m_e^*k_BT/\hbar^2)^{3/2}M_c$和 $N_v=2(2\pi m_e^*k_BT/\hbar^2)^{3/2}M_v$；这些有效状态密度是半导体电气特性评估中的重要参数。前面关于 n 和 p 的公式既适用于本征半导体（无掺杂物的半导体），也适用于掺入了施主杂质或受主杂质的半导体。掺入了杂质的相关 n 和 p 值的变化可以从简单变量——费米（Fermi）能级 E_f的变化中表现出来。

表 1-6 300K 时的 N_c和 N_v

	N_c ($\times10^{19}/\mathrm{cm}^3$)	N_v ($\times10^{19}/\mathrm{cm}^3$)		N_c ($\times10^{19}/\mathrm{cm}^3$)	N_v ($\times10^{19}/\mathrm{cm}^3$)
Ge	1.54	1.9	InSb	0.0043	0.62
Si	2.8	1.02	CdS	0.224	2.5
GaAs	0.043	0.81	CdSe	0.11	0.74
GaP	1.83	1.14	CdTe	0.13	0.55
GaSb	0.021	0.62	ZnSe	0.31	0.87
InAs	0.0056	0.62	ZnTe	0.22	0.078
InP	0.052	1.26			

来源：Böer，K. W. 1990. *Survey of Semiconductor Physics*，*Vol. 1*：*Electrons and Other Particles in Bulk Semiconductors.* Van Nostrand，New York.

n 和 p 的乘积与费米（Fermi）能级 E_f 无关，如下所示

$$n \cdot p = N_c N_v \mathrm{e}^{-E_g/k_B T}$$

式中，能隙 $E_g = E_c - E_v$。

同样，该公式既适用于本征半导体，也适用于掺杂半导体。在本征半导体中，整体电荷呈中性，因此要求 $n = p \equiv n_i$，其中 n_i 为本征载流子密度，而且

$$n_i^2 = N_c N_v \mathrm{e}^{-E_g/k_B T}$$

因此，在热平衡条件 $np \equiv n_i^2$ 下（即使是掺杂半导体），如果已知其中一种类型载流子的密度（例如，p）后，就可以直接推算出另外一种载流子的密度了（$n = n_i^2/p$）。n_i 的值随半导体材料的不同而不尽相同：CdS 为 $2 \times 10^{-3}/\mathrm{cm}^3$，GaAs 为 $3.3 \times 10^6/\mathrm{cm}^3$，Si 为 $0.9 \times 10^{10}/\mathrm{cm}^3$，Ge 为 $1.9 \times 10^{13}/\mathrm{cm}^3$，PbS 为 $9.1 \times 10^{14}/\mathrm{cm}^3$。

由于有效状态密度和温度直接相关，因此，上面的公式就不能在很广的温度范围内准确地通过 n_i 来描述温度的变化。此时，可以使用已知 N_c 和 N_v 的近似表达式：

$$n_i = 2(2\pi k_B T/\hbar^2)^{3/2}(m_e^* m_h^*)^{3/4}\sqrt{M_c M_V}\,\mathrm{e}^{-E_g/k_B T}$$

温度 $T^{3/2}$ 随 $1/T$ 的指数双倍变化，例如，在 300K 时，对于 Ge

$$n_i(T) = 1.76 \times 10^{16} T^{3/2} \mathrm{e}^{-4550T}/\mathrm{cm}^{-3}$$

对于 Si

$$n_i(T) = 3.88 \times 10^{16} T^{3/2} \mathrm{e}^{-7000T}/\mathrm{cm}^{-3}$$

5. 替位式掺杂物

本征半导体材料中只包含基本的元素原子（例如，Si 的硅原子；GaAs 的镓和砷原子等等），其电阻率非常高；因此，通过掺入特定杂质元素就可以形成可控的低电阻率材料，并形成 PN 结（半导体中 P 型区域和 N 型区域的接触面）。掺杂浓度的范围为 $10^{14} \sim 10^{17}/\mathrm{cm}^3$，比晶体中原子的密度（例如，Si 晶体中 Si 原子的密度为 $5 \times 10^{22}/\mathrm{cm}^3$）低很多。表 1-7 列出了 Si 和 GaAS 晶体中的掺杂物和它们的能级。

表 1-7　Si 和 GaAS 晶体中的受主杂质和施主杂质元素

	施主元素	$(E_c - E_d)/\mathrm{eV}$	受主元素	$(E_a - E_v)/\mathrm{eV}$
Si 晶体	Sb	0.039	B	0.045
	P	0.045	Al	0.067
	As	0.054	Ga	0.073
GaAs 晶体	S	0.006	Mg	0.028
	Se	0.006	Zn	0.031
	Te	0.03	Cd	0.035
	Si	0.058	Si	0.026

来源：Tyagi，M. S. 1991. *Introduction to Semiconductor Materials.* Wiley，New York.

图 1-6a 描述了硅晶体中受主杂质和施主杂质形成的原理。在受主杂质中，通常利用元素周期表中第Ⅲ组元素的原子来替代晶体中第Ⅳ组的硅原子。该受主原子的外层比硅原子少一个电子，因此很容易捕捉到一个自由电子来将外层共价键的价电子层（8 个电子）补充完整。受主原子因为捕捉了一个电子而带上了负电，而相邻硅原子因为外层失去一个了电子而产生了一个空穴（例如，电离时产生一个自由空穴）。由于受主杂质浓度 N_A 远大于 n_i，此时空穴的纯密度 $p \gg n_i$。基于 $np = n_i^2$ 恒量关系，电子密度 n 将在 n_i 的基础上随着 p 的增加而逐渐减小，最后就形成了 P 型半导体。在施主杂质中，通常利用元素周期表中第Ⅴ组元素的原子来替代晶体中的硅原子。相对硅原子来说，施主原子的外层存在一个多余的电子，而且该电子很容易脱离施主原子变成自由电子。这样，施主原子就变成了带正电的离子，同时还产生了一个自由电子。由于施主杂质浓度 N_D 远大于 n_i，此时电子的纯密度 $n \gg n_i$，基于 $np = n_i^2$ 恒量关系，p 将逐渐减小直至小于 n_i，最后半导体变成了 N 型半导体。

图 1-6b 给出了Ⅲ-Ⅴ型半导体中掺杂杂质的不同替代选择（GaAs 作为例

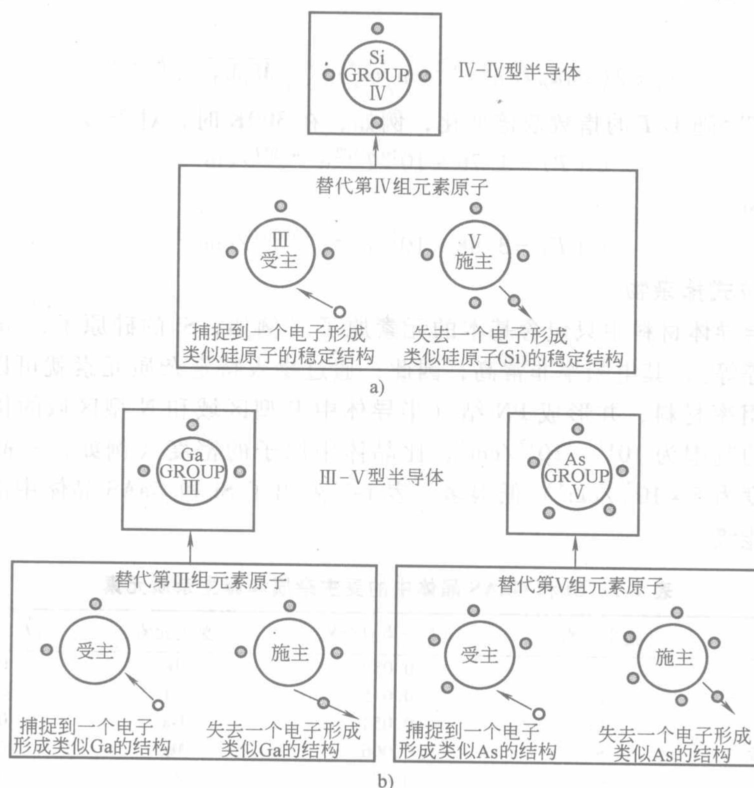


图 1-6 晶体原子的替代掺杂原子

a）Ⅳ-Ⅵ型半导体（例如，硅） b）Ⅲ-Ⅴ型半导体（例如：GaAs）

子)。用第Ⅱ组元素原子来替代第Ⅲ组元素原子，第Ⅱ组元素原子就成了受主原子（少一个电子）；而用第Ⅵ组元素原子来替代第Ⅴ组元素原子，第Ⅵ组元素原子就成了施主原子（多一个电子）。第Ⅳ组元素原子（如硅）也可以作为掺杂杂质，此时，第Ⅳ组元素原子如果替代了晶体中的第Ⅲ组元素原子，它就成了施主原子；如果替代了晶体中的第Ⅴ组元素原子，它就成了受主原子。这种在晶体中既可以作为施主杂质又可以作为受主杂质的掺杂物称为“两性杂质”。

当用来产生载流子的能量不足或很小时，可以通过受主杂质和施主杂质来产生载流子。当电离能量很小时（晶格结构中），与掺杂杂质相关的能级处于各自的能带之间（例如，施主电离能量接近于导带能级，而受主电离能量接近于价带能级）。如果电离能级和相应的价带/导带之间的级差没有超过 $3k_BT$（在 300K 时，≈0.075eV），那么在室温条件下，这些掺杂杂质基本上全部都会发生电离（称为“浅能级掺杂物”）。表 1-7 中所列掺杂物就是这类浅能级掺杂物。在掺杂了 $N_A \gg n_i$ 受主杂质的半导体中，其空穴密度 $p \approx N_A$，而少子自由电子的密度 $n \approx n_i^2/N_A$。类似的，在掺杂了 $N_D \gg n_i$ 施主杂质的半导体中，其自由电子密度 $n \approx N_D$，而少子空穴的密度 $p = n_i^2/N_D$。从以上关于费米（Fermi）能级产生载流子浓度的分析来看，费米（Fermi）能级是很容易计算的（假设已知多子载流子浓度）。

大多数半导体都可以选择性地（通过掺杂）制造成 N 型或者 P 型半导体，我们称之为“双极性半导体”。但有一些半导体只能制造成 N 型或者 P 型半导体。例如，ZnTe 只能是 P 型半导体，而 CdS 只能是 N 型半导体。这一类型的半导体，我们称之为“单极性半导体”。

能级接近于能隙中心能级的掺杂物主要作为自由电子-空穴复合的场所，同时它还影响着少子载流子的存在时间和间接能隙半导体中的主要复合机制。

1.4 载流子的转移

半导体中产生的电流既与电场作用下的自由载流子运动有关，也与载流子的扩散运动有关。在扩散运动中，载流子从高密度区域向低密度区域转移。但是，电场作用下产生的电流还是占主导地位的。

在前面的讨论中我们可以注意到，在一个理想结构的晶体中，自由电子的运动可以用一个和实际质量 m_e 相差不大的有效质量 m_e^* 来描述。这样，一旦有效质量确定了，就可以忽略理想结构晶体中的原子，其电子运动就可以看成是在自由空间中的运动。但是，如果晶体结构不是理想的，那么在忽略了理想晶格之后，那些和理想结构之间的偏差将仍然存在，并将成为晶体中电子在自由空间的散

射点。

在理想的晶体结构中，对某一元素的原子进行掺杂替代会使理想的晶体结构产生扭曲，从而导致电子发生散射。如果替代的掺杂原子发生电离，产生的电场将会加剧散射。位于间隙位置（例如，正常晶格点阵原子之间）的掺杂原子同样会破坏理想的晶体结构，并产生散射点。同样，晶体缺陷（例如，原子空位）也会破坏理想的晶体结构，并成为电子在自由空间的散射点。在实际使用的半导体晶体中，散射点的密度比硅原子的密度要小得多。这样，通过使用有效质量原子来取代硅原子就可以使散射点的空间变得很小了。

在理想的晶体结构中，所有的原子都处于晶格位置上，而且静止不动，这种情况只有在绝对零度（0K）时才会出现。当温度处于绝对零度之上时，原子就具有了热能，热能会驱使原子离开理想的位置。当原子从理想的平衡位置离开时，驱使原子回到原来位置的力量就可能产生经典的振荡器问题（原子在平衡位置振动）。这种振动会传递给相邻的原子（通过能量交换），并导致整个空间的原子都出现振动。这种如波浪般传递的振动称为“声子”，声子可以作为散射点出现在整个晶体的任何地方（即“声子散射”，也称为“点阵散射”）。

1. 低电场迁移率

硅晶体中的散射主要包括电离掺杂物散射和声子散射，而前面提到的其他散射处于次要地位。通过自由空间中的散射，自由电子在一个很高的速度（热速度 v_{therm}）上运动，该速度是由热能（$k_{\text{B}}T$）导致的。其中，$2k_{\text{B}}T/3 = 0.5m_e^* v_{\text{therm}}^2$。这样，$v_{\text{therm}} = \sqrt{4k_BT/3m_e^*}$。在室温条件的硅晶体中，热速度约为 $1.5 \times 10^7 \text{cm/s}$，该速度远远高于由电场或散射作用引起的电子运动速度。热速度和有效质量成反比关系，即有效质量较低的半导体比有效质量较高的半导体具有更高的热速度。处于较高热速度的电子在和散射中心碰撞之前都存在一个平均时间 τ_n，在这个时间里，电子的运动如同自由电子在自由空间中的运动。正是在碰撞前的这段时间内，一个外电场作用于电子并使自由电子的运行轨道产生了轻微的偏移。碰撞之后，外电场继续作用于电子，并继续对自由电子的运行轨道产生轻微的偏移。这种偏移除以平均自由时间 τ_n 就等于外电场作用下产生的速度分量。如果没有外电场的作用，电子将会朝任意的方向散射，而且不会出现偏移。在电场作用下，电子会沿电场方向产生一个漂移。因此，由电场产生的速度分量称为“漂移速度”；而此时，热速度可以忽略不计。

在标准方程式 $F = eE = m_e^* \mathrm{d}v/\mathrm{d}t$ 中，当 $v = 0$，$t = 0$ 时，再加上加速时间 τ_n，在 τ_n 时间结束后的最终速度 v_f 可以简化为 $v_f = e\tau_n E/m_e^*$；而由于 $v_{\text{drift}} = v_f/2$，因此 $v_{\text{drift}} = e\tau_n E/(2m_e^*)$。表 1-8 给出了半导体中自由电子和空穴的有效质量值。

表 1-8　常见半导体的导电性有效质量

	Ge	Si	GaAs
m_e^*	$0.12m_0$	$0.26m_0$	$0.063m_0$
m_h^*	$0.23m_0$	$0.38m_0$	0.53

外加电场 E 时，漂移速度 v_{drift} 的变化规律可以用 $v_{\text{drift}} = \mu_n E$ 来描述。其中，低电场迁移率 μ_n 的计算公式为：$\mu_n \approx e\tau_n/(2m_e^*)$。类似的，空穴也可以通过一个低电场迁移率 μ_p 来描述，$\mu_p \approx e\tau_p/(2m_h^*)$，其中 τ_p 是指空穴碰撞的平均时间。

上面的简化迁移率模型定义了一个和有效质量成反比关系的迁移率。GaAs 和 Si 晶体中的有效电子质量分别为 $0.09m_e$ 和 $0.26m_e$，由此可以看出，GaAs 的电子迁移率比 Si 高（事实上，GaAs 的电子迁移率是 Si 的 4 倍）。Si 晶体中电子和空穴的有效质量分别为 $0.26m_e$ 和 $0.38m_e$，由此可以看出，Si 晶体中的电子迁移率比空穴迁移率高（Si 晶体中，$\mu_n \approx 1400\text{cm}^2/\text{Vs}$；$\mu_p \approx 500\text{cm}^2/\text{Vs}$）。$\mu_p$ 和 μ_n 简化模型的基础是热能量导致的载流子碰撞散射条件下的简化模型。如果要得到这些迁移率的理论值，则需要更深入的分析。因此，简化模型只能作为半导体中迁移率变化的一个指导性模型，而不是预测性模型。

上面提到的 $\mu_n(\mu_p)$ 和 $\tau_n(\tau_p)$ 的线性关系为我们提供了一个关于迁移率与掺杂度和温度关系的定性理解。如前所述，在大多数半导体中，占主导地位的载流子散射是电离掺杂物散射和声子散射。而在室温和正常掺杂浓度条件下，声子散射比电离掺杂物散射更加占据主导地位。当温度下降时，晶体中原子的热能也将下降，这将导致声子散射减弱，而声子散射的平均时间将会增加。最终，根据公式 $\mu_{\text{phonon}} \approx B_1 T^{-\alpha}$，迁移率 μ_{phonon} 将随着温度的下降而增加；其中，在 300K 时，α 的范围在 1.6～2.8 之间。表 1-9 给出了 Ge、Si 和 GaAs 在室温下迁移率和温度的关系。对于电离掺杂物散射来说，越低的温度导致越低的热速度，当电子经过电离掺杂物晶格时，速度就越慢，这样散射效果就越明显。最后，迁移率 μ_{ion} 就随着温度的下降而下降。如果开始时温度很高，这时声子散射占主导地位，那么，整个阶段的迁移率一开始会随着温度的下降而逐渐提高，直到电离掺杂物散射占据主导地位后，随后迁移率才会随着温度的下降而逐渐下降。

表 1-9　300K 时，迁移率和温度的关系

	Ge		Si		GaAs	
	μ_n	μ_p	μ_n	μ_p	μ_n	μ_p
迁移率/(cm^{-2}/Vs)	3900	1900	1400	470	8000	340
温度	$T^{-1.66}$	$T^{-2.33}$	$T^{-2.5}$	$T^{-2.7}$	—	$T^{-2.3}$

来源：Wang，S. 1989. *Fundamentals of Semiconductor Theory and Device Physics*. Prentice-Hall，Englewood Cliffs，Nj.

举例来说，在 Si、Ge 和 GaAs 中，相比室温条件下的迁移率数值，温度处于 77K 时的迁移率数值只增加 7。这说明在正常的掺杂条件下，在这些半导体中占据主导地位的是声子散射。

由于产生散射概率的散射机制不同，而散射机制又定义了碰撞间的整体平均自由时间，因此不同散射机制产生的迁移率（例如，声子散射迁移率和点阵散射迁移率）就可以合并为$\mu^{-1}=\mu_{\text{phonon}}^{-1}+\mu_{\text{ion}}^{-1}$；换句话说，最小的迁移率占主导地位。

声子散射产生的迁移率与电离掺杂物的密度有关，掺杂密度越高，散射点之间的距离愈短，原子之间发生碰撞需要的时间越小。因此，温度一定时，掺杂物散射占主导地位，迁移率和掺杂水平密切相关。当电离掺杂物密度增加时，掺杂物散射决定了整体迁移率，此时温度会也升高。例如，在硅晶体中，室温条件下，$\mu_e\approx1400$，$\mu_p\approx500$ 且掺杂浓度约小于 $10^{15}/\text{cm}^3$；而当掺杂浓度大于 $10^{18}/\text{cm}^3$ 时，μ_e和 μ_p将会分别下降至 300 和 100。

以上这些定性的阐述是根据大量描述各种半导体材料的迁移率—温度和迁移率—掺杂浓度关系的细节中总结出来的，在许多参考文献中都介绍了这些内容（例如，Tyagi，1991；Nicollian and Brews，1982；Shur，1987；Sze，1981；Böer，1990；Haug，1975；Wolfe，Holonyak，and Stillman，1989；Smith，1978；Howes and Morgan，1985）[⊖]。

扩散是由载流子密度发生倾斜产生的结果。例如，扩散产生的电子流量可以由 F_e来描述，$F_e=-D_n\,\mathrm{d}n/\mathrm{d}x$；而空穴流量可以由 F_p来描述，$F_p=-D_p\,\mathrm{d}p/\mathrm{d}x$。扩散常量 D_n和 D_p与迁移率相关，公式为 $D_n=\mu_n(k_BT/e)\ \mu_e$，$D_p=\mu_p(k_BT/e)\ \mu_p$，即所谓的“爱因斯坦关系式”。具体来说就是，碰撞间的平均时间 t_{col}决定了迁移率和扩散常量。

2. 饱和载流子速度

前面讨论的迁移率是指低电场下的迁移率，因为它们只适用于场强充分低的电场情况下。低电场迁移率描述了电子从理想晶格扭曲点发生散射时的情况，电子在扭曲点发生碰撞时从电场获得了能量。在足够的电场场强情况下，电子可以获得足够的能量来和理想晶格中的原子发生非弹性碰撞。由于理想晶格中的原子密度很高（相比散射点的密度），这种新的机制决定了高电场下载流子的速度，而且该速度与场强无关（与电场强度无关，电子加速到峰值速度后就会发生非弹性碰撞，电子能量也会随之流失）。载流子速度达到饱和时的电场强度被称为“临界场强 E_{cr}”。表 1-10 归纳了 Si 晶体中自由电子和空穴的饱和速度和临界场强。GaAs 和 Ge 晶体中的饱和速度约为 6×10^6 cm/s，略低于 Si 晶体中的饱和速度。

⊖ （Tyagi，1991）是指 Tyagi 在 1991 年出版的著作，下文类似。——译者注

表 1-10　室温下 Si 的饱和速度和临界场强

	饱和速度/(cm/s)	临界场强/(V/cm)
电子	1.1×10^{7}	8×10^{3}
空穴	9.5×10^{6}	1.95×10^{4}

来源：Tyagi, M. S. 1991. *Introduction to Semiconductor Materials.* Wiley, New York.

图 1-7 给出了 Si 和 GaAs 晶体中典型的速度-电场场强特性。其中，在低电场下，迁移率和场强呈线性关系，而且 GaAs 晶体中的迁移率高于 Si 晶体。但是，Si 和 GaAs 晶体中的饱和速度 v_{sat} 没有这么大的差异。而且，与低电场迁移率不同，饱和速度与温度的关系不是很密切（因为饱和速度是由高电场强度导致的，而不是由热速度引起的）。

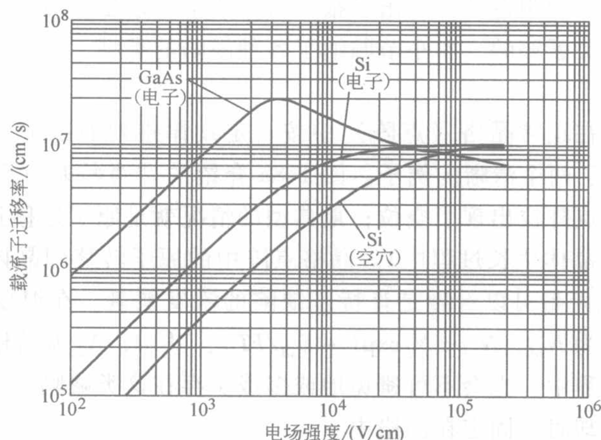


图 1-7　Si 和 GaAs 晶体的速度-电场场强特性

（来源：Sze, S. M. 1981. *Physics of Semiconductor Devices*, 2nd ed. Wiley-Interscience, New York。引用已经过允许）

图 1-7 还显示了 GaAs 晶体的一个有趣特征。随着电场强度的增加，电子速度也在逐渐增加，当超过了饱和速度值后，又会在更高的场强下逐渐回到饱和速度值。这个"逆微分迁移率"区域受到了广泛关注，并可能成为实现更高元器件运行速度的潜在途径。表 1-11 归纳了各种半导体的峰值运行速度。

表 1-11　各种半导体的饱和速度和临界场强

半导体	峰值速度/(cm/s)	半导体	峰值速度/(cm/s)
AsAs	6×10^{6}	PbTe	1.7×10^{7}
AlSb	7×10^{6}	InP	4×10^{7}
GaP	1.2×10^{7}	InSb	7×10^{7}
GaAs	2×10^{7}		

来源：Tyagi, M. S. 1991. *Introduction to Semiconductor Materials.* Wiley, New York.

当元器件的尺寸越来越小时，饱和速度就变成了更加严格的限制参数。当硅晶体的临界场强为10^4V/cm时，每1μm上1V的电压就可以产生饱和速度。在饱和速度下，电流和电压无关，不像低电场时电流和电压成比例关系。另外，对于某些半导体（例如GaAs-Si）来说，饱和速度会影响高迁移率带来的速度优势。具体来说就是，尽管低电场下越高的低电场速度会产生越高的元器件运行速度，但是饱和速度时的速度性能与高场强下的速度性能是非常相似的。

1.5 晶体缺陷

半导体晶体中存在各种各样的缺陷，其中一些会直接导致半导体性能退化；这就对晶体的生长和制造条件提出了很严格的要求，以尽量减少缺陷的发生；而其他缺陷则是无关紧要的。本小节归纳了各种缺陷的类型。

1. 点缺陷

点缺陷是指在正常晶格（点阵）中按一定规律排列的原子点阵位出现了空位，图1-8给出了两个清晰的例子。图1-8a介绍的“肖特基（Schottky）缺陷”是指在一个原子点阵位出现了空位；最典型的情况就是原子迁移到了晶体表面的正常点阵位置（晶体生长过程中）。晶格点阵中的原子转移到晶体表面时，需要一个能量E_s，该能量可以看做是肖特基缺陷的激发能量。在温度T时，肖特基缺陷的平衡密度为N_{sd}，$N_{sd}=N_L\exp(-E_{sd}/kT)$；其中，$N_L$为晶格原子密度。晶体在高温生长过程中，也会不可避免地或多或少存在这类缺陷。高温条件下的晶体缺陷在晶体冷却时会固定在晶格中。

“弗兰克（Frenkle）缺陷”是指一个原子脱离点阵位而出现在点阵位之间的

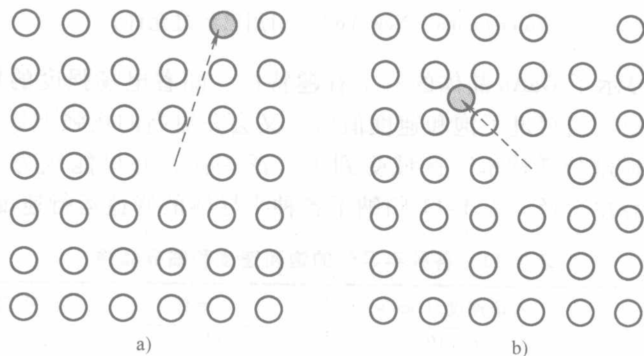


图1-8　半导体中的点缺陷

a）肖特基（Schottky）缺陷　b）弗兰克（Frenkle）缺陷

位置（例如，处于空隙位置），如图1-8b所示。因此，弗兰克缺陷对应了一对缺陷：点阵空位和处于空隙中的原子。形成这类缺陷的激发能量E_{fd}同时确定了一个非零平衡密度N_{fd}，$N_{fd}=\sqrt{N_L N_I}\exp(-E_{fd}/kT)$。同样，在晶体生长过程中也会出现弗兰克缺陷，并在晶体结晶时会固定在晶格中。

这些严重影响半导体晶体的点缺陷主要在发生电离时产生。但是，第Ⅳ组元素构成的半导体和大多数Ⅲ-Ⅴ型半导体的共价键比离子键更紧密，这就导致在产生点缺陷时需要的激发能量更大，因此在这类共价键半导体中点缺陷的密度就很小。因此，肖特基缺陷和弗兰克缺陷不会对Ⅳ-Ⅳ和Ⅲ-Ⅴ型半导体的电气性能产生很大的影响。

2. 线缺陷

线缺陷主要有3种类型（边缘位错、螺旋形位错、反相缺陷），现一一归纳如下。线缺陷是影响半导体器件电气性能的主要因素，因为它们的影响（如电子俘获中心、散射点等）会扩展到超过原子间距的地方。

边缘位错是指晶体内的原子脱离正常的规则排列，在晶体生长方向的垂直方向出现了一个多余的晶面，如图1-9a所示。这样，多余晶面附近的晶格就被破坏了，留下一个如图所示的“悬挂键”，该悬挂键会导致很多缺陷的出现。晶体中，原子在位错刚出现时就失去了一个共享的外层电子，这意味着可能会出现陷阱，并产生一连串的陷阱点，这些陷阱点之间的原子间距（远小于常见的间距）和特意掺杂的受主杂质相关。除了对晶体的电气性能产生影响外，这些缺陷还会降低晶体的机械强度。

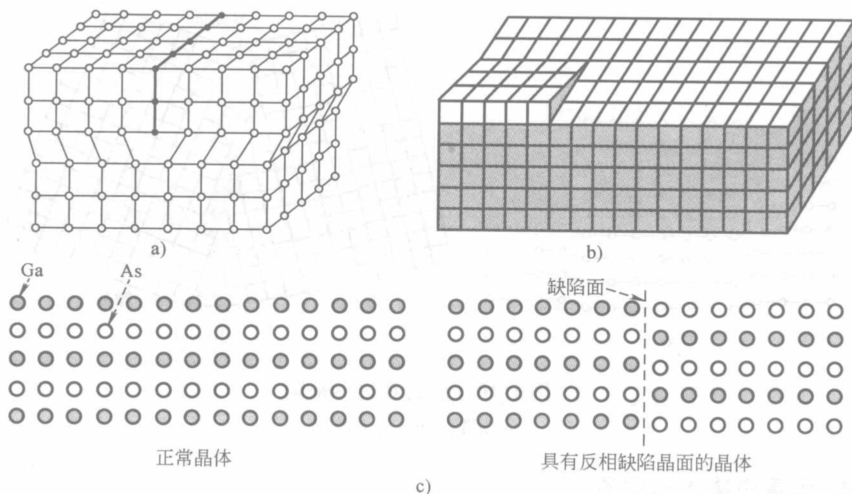


图1-9 线缺陷

a）边缘位错 b）螺旋形位错 c）反相缺陷

螺旋形位错是指晶体表面出现了一个多余的原子晶面，如图 1-9b 所示。晶体在生长过程出现了一个垂直的螺旋结构（在多余的晶面边缘）。另外，在晶体生长过程中，扩展区域中晶格结构的变化会导致很多晶体腐蚀特性发生变化，如结退化等等。另外，悬挂键也会作为陷阱影响晶体的电气特性。

反相缺陷只出现在化合物半导体中。图 1-9c 给出了Ⅲ-Ⅴ型半导体晶体的一部分，其中相比左侧的规则排列，右侧相互交错的 Ga 和 As 原子层是相互反相的。这种相位上的差错直接导致了原子在晶面两侧相互结合上的缺陷，从而影响了元器件的电气特性。克服这类缺陷的做法是，在大直径晶体生长过程中要采取精密的制造工艺，以确保整个晶面同时形成。

3. 层错和边界缺陷

层错是指在大晶体中出现了一个多余的小面积原子晶面。层错直接造成正常的晶面被扭曲并扩展以容纳多余的晶面（见图 1-10a 所示）。在图 1-10a 中，在正常的晶面位置上出现了一大块层错晶面。这些缺陷严重影响了晶体的电气特性。

边界缺陷是指两个正在生长的、具有不同结晶方向的微晶体在结合时，在两个晶格的晶面分隔处留下了一条缺陷，如图 1-10b 所示。从图中可以看出，两个结晶方向出现偏差的晶体其中间空隙已经被各自的晶格填满，从而留下了一个缺陷面。边界缺陷影响了边界处的电阻率，并导致了陷阱以及其他电气效应的产生。在多晶硅（细小的晶体硅颗粒）中，边界缺陷带来的影响包括硅晶体中的颗粒效应以及颗粒边缘区域效应（如同晶格间的互联网）。

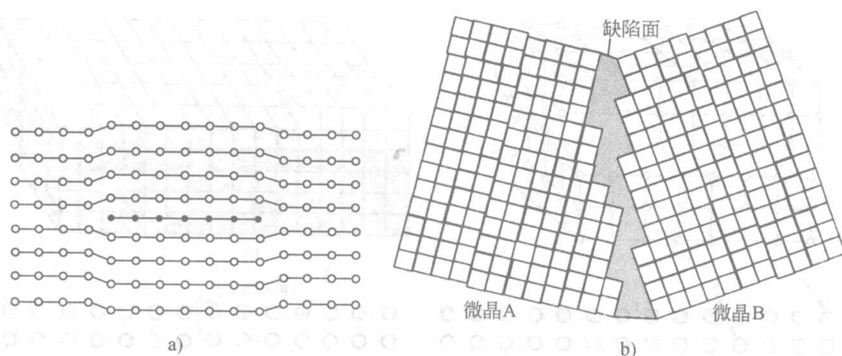


图 1-10　层错和边界缺陷

a）层错　b）边界缺陷

4. 无意中掺入的杂质

由于化学制品和环境等因素的影响，在晶体生长和元器件生产过程中，难免会有一小部分的杂质混进了半导体晶体当中，我们称之为“无意中掺入的杂

质”。其中，那些取代晶体中原子点阵位置的杂质称为“替代式杂质”；其他处于点阵位置之间的杂质称为“间隙杂质”。其中，有一些杂质不会对晶体的电气特性产生严重影响；还有一些杂质则会对晶体的电气特性产生有益影响。例如，可以补偿悬挂键的氢元素，以及那些提高深度能量俘获等级的元素（例如，能级接近能隙中心水平）。这种深度能量俘获等级对于间接能隙半导体复合时间常量的确定具有很重要的意义。事实上，金元素很早就用在了双极性硅晶体管中，用来提高载流子复合率并提升晶体管的运行速度。除了这些良性杂质外，其他杂质则会对元器件的性能产生有害的影响。

光电元器件对那些无意中掺入的杂质很敏感，许多电气性能陷阱能级都和缺陷相关。表 1-12 归纳了一些在 GaAs 中确定的陷阱能级，其中一些是由无意中掺入的杂质引起的，而另一些则是由缺陷引起的。因此，半导体晶体的生长过程和材料生产工艺是一个非常复杂的策略性过程。在这个过程中必须尽量减少无意中掺入杂质的发生，并在一个合适的水平选择性地添加一些有益的杂质，另外还要对那些无意中掺入的良性杂质采取有效的保留策略。

表 1-12 GaAs 中的陷阱类型（能量为 E_t）

类型	E_c-E_t	名称	类型	E_t-E_v	名称
浅施主杂质	≈5.8meV		浅受主杂质	≈10meV	
施氧杂质	0.82eV	EL2	受锡杂质	0.17eV	
受铬杂质	0.61eV	EL1	受铜杂质	0.42eV	
深受主杂质	0.58eV	EL3	空穴陷阱	0.71eV	HB2
电子陷阱	0.90eV	EB3	空穴陷阱	0.29eV	HB6
电子陷阱	0.41eV	EB6	空穴陷阱	0.15eV	HC1

来源：Shur，M. 1987. *GaAs Devices and Circuits*. Plenum Press，New York.

5. 面缺陷：重构面

假设沿着结晶面的方向将单晶体切成一片一片的晶面，那么两片晶面之间的高密度原子键就被破坏了。如果晶体结构是完全向着晶体表面的方向扩展的，那么悬挂键就可以用来描述晶体的表面结构。但是，原子晶格和由于切片损坏的键并不能描述晶体表面的最小能级状态。因此，一个重新构建的原子键和原子点阵在晶体表面形成了晶体的最小能量结构，这个过程称为“表面重构”，表面重构在晶体表面形成了不同的晶格结构。表面重构过程对很多条件都很敏感，这使得实际半导体晶体中的晶体表面变得非常复杂。特别重要的是那些可以与半导体基本原子不同类型原子混合的晶体表面层（例如，硅半导体中的本地氧化物）。晶体表面的重要性在硅 MOS 晶体管中可以很清楚地体会到，目前在 MOS 晶体管的界面已经可以实现很低的缺陷率了。重构的晶体表面会很明显地影响元器件的电气特性。例如，晶体表面区域载流子的迁移率相比整体的迁移率会大大降低。另外，MOS 器件对固定电荷和陷阱点阵非常敏感，这些陷阱点阵会降低半导体的表面电压，该表面电压和实际的门电压相关。

表面重构是一个非常复杂的过程，也是一个很具体的主题，在此不作详细介绍。实际上，目前的微制造技术已经发展到了能实现高质量晶体表面的水平。

1.6 小结

多种多样的半导体材料已经被广泛应用到了很多实际需求中，这些材料通过非常先进的元器件制造工艺变成了硅微电子元器件和Ⅲ-Ⅴ型光电子元器件。至少在下一个十年中，大部分的计算机、信息技术和消费类电子产品的革新必需依靠这些重要的半导体材料。当我们使用这些半导体材料（如，硅）来制造微小元器件时，一些新的半导体材料也在不断涌现，而且它们可能具有更高的速度、更低的功耗以及其他一些优势。本章给出了关于半导体材料的一个总体概述。参考文献列表中有许多很好的、具有很强可读性的书籍，它们介绍了本章中没有的涉及到的更深的相关知识。表 1-13 给出贯穿本章的一些相关信息列表。

表 1-13 室温下，GaAs 和 Si，CdS 和 CdSe 半导体的特性既呈现闪锌矿形式，又呈现纤维锌矿形式

	晶格类型	晶格常数 /Å	能隙 /eV	电子有效质量	空穴有效质量	介电常数	电子迁移率 /(cm^2/V - s)	空穴迁移率 /(cm^2/V - s)
C	金刚石	3.5668	5.47(I)①	0.2	0.25	5.7	1800	1200
Si	金刚石	5.4310	1.11(I)	m_l:0.98 m_t:0.19	m_l:0.16 m_h:0.5	11.7	1350	480
Ge	金刚石	5.6461	0.67(I)	m_l:1.58 m_t:0.08	m_l:0.04 m_h:0.3	16.3	3900	1900
AlP	闪锌矿	5.4625	2.43	0.13		9.8	80	
AlAs	闪锌矿	5.6605	2.16(I)	0.5	m_l:0.49 m_h:1.06	12.0	1000	180
AlSb	闪锌矿	6.1355	1.52(I)	0.11	m_l:0.39	11	900	400
GaN	纤维锌矿	3.189	3.4	0.2	0.8	12	300	
GaP	闪锌矿	5.4506	2.26(I)	0.13	0.67	10	300	150
GaAs	闪锌矿	5.6535	1.43(D)②	0.067	m_l:0.12 m_h:0.5	12.5	8500	400
GaSb	闪锌矿	6.0954	0.72(D)	0.045	0.39	15.0	5000	1000
InAs	闪锌矿	6.0584	0.36(D)	0.028	0.33	12.5	22600	200
InSb	闪锌矿	6.4788	0.18(D)	0.013	0.18	18	100000	1700
InP	闪锌矿	5.8687	1.35(D)	0.077	m_l:0.12 m_h:0.60	12.1	4000	600
CdS	闪锌矿	5.83	2.42(D)	0.2	0.7	5.4	340	50
CdSe	闪锌矿	6.05	1.73(D)	0.13	0.4	10.0	800	
CdTe	闪锌矿	6.4816	1.50(D)	0.11	0.35	10.2	1050	100
PbS	NaCl	5.936	0.37(I)	0.1	0.1	17.0	500	600
PbSe	NaCl	6.147	0.26(I)	m_l:0.07 m_t:0.039	m_l:0.06 m_h:0.03	23.6	1800	930
PbTe	NaCl	6.45	0.29(I)	m_l:0.24 m_t:0.02	m_l:0.3 m_h:0.02	30	6000	4100

① 间接能带。

② 直接能带。

来源：Wolf, C. M., Holonyak, N., and Stillman, G. E. 1989. *Physical Properties of Semiconductors*, Prentice-Hall, New York.

名词解释

双极性半导体：可以通过选择性地掺入微量杂质来实现 N 型或 P 型材料的半导体。

两性杂质：既可以作为施主杂质又可以作为受主杂质的掺杂物。

能隙：导带和价带之间的能量差。

化合物半导体：由元素周期表中不同组的两种或两种以上元素原子构成的半导体。

导带：半导体晶体中自由电子处于静止状态时的能级。

深能级杂质：能级处于能隙中间的掺杂物或杂质。在间接能隙半导体中对载流子的复合起重要作用。

直接能隙半导体：导带最小值和价带最小值在相同的波矢量出现（相同的时机），这一点对于光学材料非常重要。

有效质量：半导体中载流子运动时的质量等价值，如同在自由空间中运动时的等价状态。

元素半导体：晶体由单一原子元素组成的半导体。

间接能隙半导体：导带最小值和价带最小值在不同的波矢量出现（不同的时机）。

本征半导体：半导体晶体中没有任何无意中掺入的杂质或特意掺入的杂质（掺杂物）。

低电场迁移率：载流子速度与低电场场强之间的对应关系常量。

四元半导体：半导体中包含 4 种元素，其中两种来自元素周期表的同一组，另外两种来自周期表的另一组。

饱和速度：由电场作用导致的最大载流子速度（由无弹性扩散引起）。

浅能级杂质：能级非常接近施主杂质导带或受主杂质价带的掺杂物。

替代掺杂物：晶体中替代正常原子晶格位置的掺杂物。

三元半导体：半导体中包含 3 种元素，其中两种来自元素周期表的同一组，另外一种来自周期表的另一组。

单极性半导体：通过选择性掺杂只能形成 N 型或 P 型材料的半导体。

价带：半导体晶体中价电子和自由空穴处于静止时的能级。

参考文献

[1] Beadle, W. E., Tsai, J. C. C., and Plummer, R. D., eds., 1985. *Quick Reference Manual for Silicon Integrated Circuit Technology*. Wiley, New York.

[2] Böer, K. W. 1990. *Survey of Semiconductor Physics, Vol. 1: Electrons and Other Particles in Bulk Semiconductors.* Van Nostrand, New York.

[3] Böer, K. W. 1992. *Survey of Semiconductor Physics, Vol. 2: Barriers, Junctions, Surface, and Devices.* Van Nostrand, New York.

[4] Capasso, F. and Margaritondo, G., eds. 1987. *Heterojunction and Band Discontinuities.* North Holland, Amsterdam.

[5] Haug, A. 1975. *Theoretical Solid State Physics, Vols. 1 and 2.* Pergamon Press, Oxford, England.

[6] Howes, M. J. and Morgan, D. V. 1985. Gallium Arsenide: Materials, Devices, and Circuits. Wiley, New York.

[7] Irvine, S. J. C., Lum, B., Mullin, J. B. and Woods, J., eds. 1982. Ⅱ-Ⅵ *Compounds* 1982. North Holland, Amsterdam.

[8] Lannoo, M. and Bourgoin, J. 1981. *Point Defects in Semiconductors.* Springer-Verlag, Berlin. Loewrro, M. H., ed. 1985. *Dislocations and Properties of Real Materials.* Inst. of Metals, London.

[9] Moss, T. S. and Balkanski, M., eds. 1980. *Handbook on Semiconductors, Vols. 2: Optical Properties of Solids.* North Holland, Amsterdam.

[10] Moss, T. S. and Keller, S. P., eds. 1980. *Handbook on Semiconductors, Vols. 3: Material Preparation.* North Holland, Amsterdam.

[11] Moss, T. S. and Paul, W., eds. 1982. *Handbook on Semiconductors, Vols. 1: Band Theory and Transport Properties.* North Holland, Amsterdam.

[12] Nicollian, E. H. and Brews, J. R. 1982. *MOS Physics and Technology.* Wiley, New York.

[13] Pantelides, S. T. 1986. *Deep Centers in Semiconductors.* Gordon&Breach Science, New York.

[14] Shur, M. 1987. *GaAs Devices and Circuits.* Plenum Press, New York.

[15] Smith, R. A. 1978. *Semiconductors.* Cambridge University Press, London.

[16] Sze S. M. 1981. *Physics of Semiconductors Devices*, 2nd ed. Wiley Interscience, New York.

[17] Tyagi, M. S. 1991. *Introduction to Semiconductor Materials.* Wiley, New York.

[18] Wang, S. 1989. *Fundamentals of Semiconductor Theory and Device Physics.* Prentice-Hall, Englewood Cliffs, New Jersey.

[19] Willardson, A. K. and Beer, A. C., eds. 1985. *Semiconductors and Semimetals, Vol. 22.* Academic Press, New York.

[20] Wilmsen, C. W., ed. 1985a. *Physics and Chemistry of* Ⅲ-Ⅴ *Compounds.* Plenum Press, New York.

[21] Wilmsen, C. W., ed. 1985b. *Physics and Chemistry of* Ⅲ-Ⅴ *Compound Semiconductor Interface.* Plenum Press, New York.

[22] Wolfe, C. M., Holonyak, N., and Stillman, G. E. 1989. *Physical Properties of Semiconductors.* Prentice-Hall, New York.

备注

参考文献列表中的书籍都是该领域很受欢迎的书籍，对半导体材料知识想有更进一步了解的读者可以阅读这些书籍。电气和电子工程师学会（IEEE, Piscataway, NJ）出版了几期关于最新半导体材料和元器件应用的期刊，包括《*IEEE Trans. Electron Devices*》、《*IEEE Solid State Circuits Journal*》、《*IEEE Journal on Quantum Electronics*》和《*IEEE Journal on Lightwave Technology*》。美国物理协会（American Physical Society）

也出版了几期关于半导体材料的期刊。其中之一是关于材料持续研究的，刊名为《*Journal of Applied Physics*》。美国光学协会（The Optical Society of America）出版了期刊《*Applied Optics*》，该期刊涵盖很多实际的要点和最新导向。光学照相仪器工程师协会（The Society of Photo-Optical Instrumentation Engineers，SPIE，Bellingham，WA）主办了一系列关于光电子材料和其他材料的会议。感兴趣的读者可以直接和 SPIE 联系，索取会议学报。

第2章　热效应特性

David F. Besch

2.1　引言

电子和电气元器件性能的额定值和该元器件的散热能力有很大的关系。随着元器件微型化的发展，工程师们越来越关注的是散热能力以及元器件在性能和物质结构上的变化与温度的关系。下面将集中介绍热效应原理。电子元器件材料的分类方式是多种多样的，本章中的电子元器件材料主要是根据它们的电阻率来分类：

1）绝缘体。

2）半导体。

3）导体。

从这个细致的分类中我们可以看出，有些材料不只适合一种分类。例如，陶瓷是绝缘体，但是在混合一些其他元素之后，就可以作为半导体、晶体管甚至导体了。一般来说，虽然材料电阻率和温度之间的关系很重要，但是设计工程师更加关心的是电阻值随着温度变化多少以及这种变化关系是否会导致电路参数不符合技术要求。

2.2　热效应基本原理

在物质的常用模式中，热量是一种与物质分子、原子、离子的位置和活动相关的能量形式。其中，物质中粒子的位置情况相当于物质的状态，对应的是一种潜在能量即势能，而物质分子、原子、离子的活动对应的则是动能。加热可以使物质温度升高，反之亦然。加热也可以使固体转化为液体，使液体转化为气体，这两种过程是物质状态形式发生变化的典型情况。热量是通过卡路里（cal）、英国热量单元（Btu）或焦耳（J）来计量的。1卡路里相当于将1g水的温度提高1℃（如14.5～15.5℃）时所需要的能量；1Btu相当于将1磅[1]（1lb）水的温度提高1华氏摄氏度[2]（1℉）时所需要的能量；而1J相当于1N的力使物体移

---

[1] 1磅（1lb）=0.453kg。

[2] 1华氏摄氏度（℉）$=\frac{5}{9}$K$=\frac{5}{9}$℃。

动 1m 的距离所做的功。因此，热能可以转化成机械能来做功。3 种计量单位之间的比例关系为：1Btu = 251.996cal = 1054.8J。

1. 温度

温度是物质平均动能的一种表现形式，也可以看做是物质之间热含量差异的计量方式。温度的计量方式可以是华氏温度比例，也可以是摄氏温度比例。华氏温度将水的冰点温度计为 32℉，将水的沸点温度计为 212℉；而摄氏温度或者称为“百分度”温度将水的冰点温度计为 0℃，将水的沸点温度计为 100℃。

Ranking 比例是基于华氏温度的一种绝对温度比例，而 Kevin（开尔文）比例是基于摄氏温度的一种热力学温度比例。绝对温度是指在 0 气压时氢温度计上对应的 0 度。从上面关于温度的定义看，0°R 和 0K 对应的是 0 动能。

以上 4 种温度比例之间的关系如下：

℃ = 5/9(℉ − 32)

℉ = 9/5(℃) + 32

K = ℃ + 273.16

°R = ℉ + 459.69

2. 热量

热量的定义为，在不改变物质状态的前提下，将 1mol 的物质原子温度提高 1℃所需的能量。从热量的定义可以看出，热量可以用来衡量单位质量的物质随温度变化所产生的能量变化。热量是物质的基本特性之一，计量单位为每℃ cal/g 或每℉ Btu/lb，$c_p = \frac{\partial H}{\partial T}$。

3. 比热容

比热容是指一种物质的热量相对参考物质（通常为水）热量的比例关系。由于水的热量为 1Btu/lb 和 1cal/g，因此在数值上，比热容就等于热量。

4. 热导性

热能通过物质的热导性进行传递，原子和离子在振动时以较低的能量进行热能传递。能量流是由自由电子引起的，公式为

$$Q = kA \frac{\partial T}{\partial l}$$

式中，Q 为单位时间的热流量；k 为热导率；A 为热路径面积；l 为热路径长度；T 为温度。

热导率 k 与温度密切相关，当温度超过室温时，热导率 k 值将下降。

5. 热膨胀

当物质被加热时，物质中的原子和离子动能会随之增加。换句话说，加热使物质超出常温状态并产生了与温度变化成比例关系的热膨胀。如果物质在加热或降温的过程中无法膨胀或缩小，那么物质就会产生内部压力。公式为

$$\frac{\partial l}{\partial T}=\beta_L l \text{ 和 } \frac{\partial V}{\partial T}=\beta_V V$$

式中，l为长度；V为体积；T为温度；β_L为线性膨胀系数；β_V为体积膨胀系数。

6. 固体

固体是物质的一种状态形式，处于该状态的物质其原子或离子之间的引力产生的势能远远大于原子或离子振动的动能。原子之间的相互吸引力使得大多数物质都形成了晶体结构。非晶体固态物质称为“非晶态物质”，包括玻璃、大多数塑料以及一些经过冷却从液态迅速凝固的半固态金属。非晶态物质没有广泛的规则原子排列。

晶体物质可以结晶成以下几种几何模型：

1）立方体。

2）四边形。

3）正交（晶）。

4）单晶。

5）六边形。

6）菱形。

通常，物质的特性是指与晶体晶面的密度和生长方向相关的一些功能。

一些物质在仍然是固体时其状态已经悄悄发生了变化。当这些物质加热时，纯离子会在温度达到912℃时，从晶体的晶格中心转移到晶体表面，而晶体中的原子间距也会由于热膨胀从0.12nm扩展到0.129nm。那些由成分相同的两种或两种以上不同类型晶体组成的物质称为“多形态物质”。

7. 液体

液体也是物质的一种状态形式，处于该状态的物质原子或离子之间的引力产生的势能近似等于原子或离子振动的动能。液体在自身重力作用下流动。物质从固体转化成液体的过程称为“熔解”。物质熔解时需要一个特有的能量，称为“熔解热量”。物质在熔解时，原子晶体会经历一个扭曲变形的过程，最后大多数物质都会出现体积膨胀的现象。在某些物质中（如水），其立体定向的共价键和较少填充因子的特点使得在加热过程中仍能保持紧密的物质结构。

8. 气体

气体也是物质的一种状态形式，处于该状态的物质原子或离子振动的动能远远大于原子或离子之间的引力产生的势能。在压力一定时，气体的体积会随着温度的变化而呈正比变化；在体积一定时，气体的压力也会随着温度的变化而呈正比变化；在温度一定时，给定质量的气体的体积会随着压力的变化而呈反比变化。这三者之间的关系可以通过气体规律公式体现：

$$PV=RT$$

式中，P 为绝对压力；V 为指定的体积；R 为通用气体常量 t；T 为绝对温度。

物质从液体转化成气体时需要一个特有的能量，称为“汽化热量”。

2.3　其他材料特性

1. 绝缘体

绝缘体是指电阻率远大于 $10^7\Omega cm$ 的材料。大部分陶瓷制品、塑料、各种氧化物、纸制品和空气都是绝缘体。氧化铝（Al_2O_3）和氧化铍（BeO）等陶瓷制品通常用作半导体工艺中的衬底或芯片载体。一些陶瓷和塑料制成的薄膜也可以用作电容中的绝缘体。

2. 绝缘常数

电容中的两个导电板块通过一个绝缘体薄膜相互隔离。电容容量和绝缘材料的绝缘常数呈正比关系。在陶瓷化合物中掺入钛酸钡后就具有了很高的绝缘常数，可以用于电容器中；塑料制品如云母、聚苯乙烯、聚碳酸酯和聚酯薄膜也可用于电容器中。电容容量会随着温度的变化而变化。参考 2.4 节的第一部分可以看到不同温度时电容容量的计算方法。

3. 电阻率

绝缘体的电阻率会随着温度的升高而逐渐下降。

4. 半导体

半导体是指电阻率在 $10^{-4}\sim10^7\Omega cm$ 范围的材料。硅（Si）、锗（Ge）和 GaAs 就是典型的半导体材料，其电阻率以及相反对应的电导率也各不相同。在 Si 和 Ge 本征半导体中，对于不同的掺杂元素来说，电导率和温度之间的关系可以用如下公式来描述

$$\sigma=\sigma_0 e^{\frac{E_g}{2kT}}$$

式中，σ 为电导率；σ_0 为常量 t；E_g 为 1.1eV（Si）；k 为波尔兹曼常量 t；T 为温度，单位为 K。这样，当温度从 27K 升高到 200K 时，Si 的电导率会增加 2400。

5. 导体

导体是指电阻率小于 $10^{-4}\Omega cm$ 的材料，包括金属、金属氧化物以及导电型非金属。导体的电阻率会随着温度的升高而增加，如图 2-1 所示。

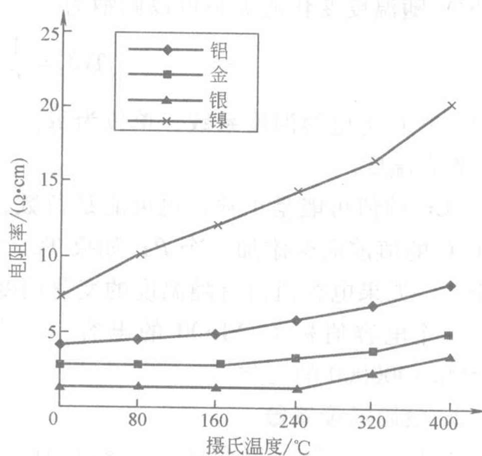


图 2-1　电阻率与温度的关系

6. 熔点

焊接材料是一种广泛应用在电子系统中的重要物质。锡制焊料是最常用的焊接化合物。在系统平衡图中说明了当锡含量为61.9%时焊接材料最容易熔化（即形成低共熔混合物）；因此，常用的焊接材料都是合金产品。高铅含量焊接材料中锡的含量为10%，常用作高温焊接材料。而高锡含量的焊接材料常用于特殊用途，如高腐蚀性环境。表2-1列出了部分常用的合金。

表2-1　部分合金的特性

锡(%)	铅(%)	银(%)	温度/℃
60	40	—	190
60	38	2	192
10	90	—	302
90	10	—	213
95	5	5	230

2.4 工程数据

从电阻率和绝缘常数与温度的变化关系曲线图中很难得出电子元器件性能的具体参数值。电子设计工程师更加关心的是电阻率如何随着温度变化以及这种变化关系是否导致电路参数不合规范。接下来我们将给出常用元器件在温度变化时的参数定义。

1. 电容温度系数

由于绝缘常数随温度变化，因此电容容量也随温度变化。电容温度系数（TCC）随温度变化的关系可以归纳为

$$\mathrm{TCC}=\frac{1}{C}\frac{\partial C}{\partial T}$$

式中，TCC为电容温度系数，单位为每摄氏度百万分之（ppm/℃）；C为电容容量；T为温度。

TCC的值可能是正数，也可能是负数，还可能是0。如果TCC的值是正数，在TCC的值前面会添加一个P；如果TCC的值是负数，在TCC的值前面会添加一个N。如果电容值没有随温度的变化而变化，那么电容就会被标记为NPO。例如，一个电容值标为$N1500$的电容，在温度每变化1℃时，其电容值会产生$-1500/1000000$的变化。

2. 电阻温度系数

由于电阻率随温度变化，因此电阻的值也会随着温度而变化。电阻温度系数（TCR）用来描述这一变化关系，公式为

$$\mathrm{TCR}=\frac{1}{R}\frac{\partial R}{\partial T}$$

式中，TCR 为电容温度系数，单位为每摄氏度百万分之（ppm/℃）；R 为电阻值；T 为温度。

TCR 的值可能是正数，也可能是负数，还可能是 0。表 2-2 给出了常用电阻的 TCR 值。表 2-2 中最后 3 个是嵌入在单片式硅集成电路中电阻的 TCR 值。

表 2-2　常见电阻的 TCR 值

电阻类型	TCR/(ppm/℃)		
碳含量	+500	to	+2000
绕线	+200	to	+500
厚膜	+20	to	+200
薄膜	+20	to	+100
扩散基极	+1500	to	+2000
扩散发射极	+600		
注入离子	±100		

3. 温度补偿

温度补偿是指电子设计工程师为提高电路或系统性能和稳定性，尽量减少温度变化带来的影响而采取的措施的总称。除了使用最合适的电容 TCC 值和电阻 TCR 值之外，还通常采用以下的元器件和技术。

1）热敏电阻。

2）电路设计稳定性分析。

3）热能分析。

4. 热敏电阻

热敏电阻是一种电阻值变化范围很大的半导体电阻。它在正负温度系数下均可使用，并主要用于系统温度测量和控制，以及温度补偿。在用于温度补偿时，主要用来补偿由于温度变化带来的不必要的电阻变化。

5. 电路分析

在由半导体元器件构成的模拟电路中，存在一些由于温度变化导致的电路偏置稳定性隐患。通过结型元器件的电流呈指数变化关系，公式如下

$$i_D=I_S\left(\mathrm{e}^{\frac{qV_D}{nkT}}-1\right)$$

式中，i_D为流过 PN 结的电流；I_S为反向饱和电流；V_D为 PN 结两端的电压降；q 为电子电荷量；n 为扩散系数；k 为玻耳兹曼常数；T 为热力学温度。

结型二极管和双极性结型晶体管中的电流呈上述指数变化关系。一些偏置电路相比其他电路具有更好的温度稳定性。电路设计师可以通过电路的局部温度系数 TC_F来评估电路性能，TC_F由下式计算：

$$TC_F = \frac{1}{v(T)} \frac{\partial v(T)}{\partial T}$$

式中，$v(T)$ 为电路变量；TC_F为温度系数；T 为温度。

通常，商用电路的仿真程序可以用来评估给定电路的性能随温度变化的结果。例如，SPICE 可以在任何温度下通过非常精密的模型来实现对所有电路元器件的仿真。

6. 热能分析

那些体积很小或功耗很大的电子系统的内部温度在工作时会随着时间逐渐上升。热能分析就是设计师用来评估元器件中热量转移过程的技术，在该转移过程中，热量从本器件中转移到周围环境中。

名词解释

低共熔混合物：在两种合金的熔解度曲线的交集中取最低溶解温度值的合金。

立体定向：两个原子之间的定向共价键。

参考文献

[1] Guy, A, G. 1967. *Element of Physical Metallurgy*, 2nd ed. , pp. 255-276. Addision-Wesley, Reading, MA.

[2] Incropera, F. P. and Dewitt, D. P. 1990. *Fundamentals of Heat and Mass Transfer*, 3rd ed. , pp. 44-46. Wiley, New York.

备注

关于半导体材料热效应特性的知识可以参考以下资料：

[1] Banzhaf, W. 1990. *Computer-Aided Circuit Analysis Using·Psice.* Prentice-Hall, Englewood Cliffs, NJ.

[2] Smith, W. F. 1990. *Principles of Material Science and Engineering.* McGraw-Hill, New York.

[3] Van Vlack, L. H. 1980. *Elements of Material Science and Engineering.* Addison-Wesley, Reading, MA.

第3章 半 导 体

Sidney Soclof

3.1 引言

晶体管构成了当今电子元器件和电子系统的基础，如广泛应用在收音机、电视机以及计算机中的各种集成电路。晶体管是一种固体电子元器件，由一种称为半导体的材料制造而成。尽管砷化镓（GaAs）作为一种化合物半导体材料可以用来制造一些高速的晶体管，但是到目前为止，在制造晶体管时应用最广泛的半导体材料仍然是硅。

半导体

半导体是指电导率介于导体和绝缘体之间的物质。良好的导体包括所有的金属，其电阻率低于$10^{-6}\Omega\cdot cm$；绝缘体的电阻率处于$10^{6}\sim10^{12}\Omega\cdot cm$范围；而半导体的电阻率通常在$10^{-4}\sim10^{4}\Omega\cdot cm$范围。半导体的电阻率与其中的杂质有很大关系，这种杂质称为“掺杂物”，是特意添加到半导体中用来改变其电子特性的一种物质。

我们首先来介绍纯净半导体或本征半导体。在半导体材料中，由于热能作用，共价键中的电子可以摆脱束缚成为自由电子，自由电子在半导体中自由移动并成为半导体导电时的电荷载流子。电子在摆脱共价键的束缚后留下的空位称为“空穴”，相邻共价键中的电子很容易移动到该空位或空穴中，这样该电子在原来的共价键中就留下来一个空穴，这个过程看上去像是空穴从一个共价键中转移到了另一个共价键中。随着这个过程的不断继续，我们可以认为空穴是在半导体中移动。这些空穴相当于带了和电子电荷等量的正电荷，也成为半导体导电时的电荷载流子。因此，在半导体中存在两种用于导电的电荷载流子，即自由电子和空穴。由于自由电子和空穴同时产生，又同时复合，因此，在本征半导体中，自由电子和空穴的数量相同。

在非本征或掺杂半导体中，掺入的杂质就是专门用于调节导电性能的。例如，在硅半导体中，每个硅原子通过共价键和相邻的4个硅原子共享价电子层中的4个价电子。如果掺入的杂质原子具有5个价电子，如磷原子，该原子替代了硅原子，那么其5个价电子中的4个价电子会和周围的硅原子形成共价键，而多余的一个电子就没有共价键的束缚了。在常温条件下，该电子就会摆脱原子的束

缚成为自由电子。这种为半导体带来自由电子的五价掺杂物称为“施主杂质”。这些由施主杂质带来的自由电子打破了原先自由电子和空穴之间的数量平衡，使得半导体中自由电子的数量远大于空穴的数量，这种半导体称为“N 型半导体”，其中，自由电子是多数载流子，而空穴是少数载流子。在 N 型半导体中，自由电子浓度比空穴浓度高出很多个数量级。

如果在硅半导体中掺入的杂质原子只有3个价电子，如硼原子，该原子替代了硅原子，那么硼原子的 3 个价电子和周围的硅原子形成共价键，而其中一个共价键就缺少了一个共享电子，形成了一个电子空位，即空穴。这时，相邻的硅原子间形成的共价键中的电子很容易跳转到该电子空位，并在原来位置形成了另一个电子空位，即空穴。因此，这种三价掺杂原子接受了自由电子并产生了空穴，称为“受主杂质”。这些由受主杂质产生的空穴打破了原先自由电子和空穴之间的数量平衡，使得半导体中空穴的数量远大于自由电子的数量，这种半导体称为“P 型半导体”。在 P 型半导体中，空穴浓度比自由电子浓度高出很多个数量级。

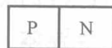


图 3-1 PN 结

图 3-1 给出了在单晶体硅中 PN 结的例子。在左侧晶体中，掺入了受主杂质形成了 P 型半导体；在右侧晶体中，掺入了施主杂质形成了 N 型半导体。在两块晶体的接合处就形成了“PN 结”。由于两侧的自由电子和空穴浓度相差很大，因此自由电子和空穴将会产生扩散并穿过两块晶体的接合处，到达另一侧。最后，N 型半导体这一侧就获得了来自 P 型半导体一侧的正电荷。这样，在接合处就形成了势垒电压。这种势垒电压的最大值称为“接触电压”。电压势垒的形成过程如下：基于平衡的条件，P 型侧的多数载流子空穴的浓度由于扩散到达 N 型侧后（该过程有助于提高势垒电压）将逐渐减小，直到和 P 型侧的少数载流子空穴扩散到 N 型侧（该过程有助于减小势垒电压）后空穴浓度相当。类似的，N 型侧的多数载流子自由电子浓度由于扩散将逐渐减少直到和 P 型侧的少数载流子自由电子浓度相当。这样，穿过接合处的净电流在动态平衡的条件下仍为 0。

3.2 二极管

在图 3-2 的电路中，硅片相当于一个二极管或者双接线端的电子元器件，图中给出了在两端添加偏置电压的情形。在图 3-2a 中，偏置电压为正向电压。在正向电压的作用下，PN 结中的势垒电压会减小，而且穿过 PN 结的自由电子和空穴的数量会急剧增加，形成的正向电流会增加大约一个指数级。正向电流的值随正向电压变化的关系如下面的公式所示：

$$I = I_0(\exp(qV/nkT) - 1)$$

式中，I_0 为反向饱和电流；q 为电子电量；n 为 1 和 2 之间的无量纲因数；k 为

波尔兹曼常数；T 为热力学温度，单位为 K。

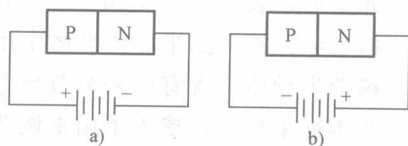


图 3-2 二极管偏置

a）正向偏置 b）反向偏置

如果将热电压定义为 $V_T = kT/q$，二极管的电流公式可以写成：

$$I = I_0(\exp(V/nV_T) - 1)$$

在室温下，$V_T \cong 26\text{mV}$。对于硅晶体二极管，n 约为1.5。

在图3-2b 中，偏置电压为反向电压。在反向电压的作用下，PN 结中的势垒电压会增加，并彻底阻止了 PN 结两侧自由电子向 P 型区域的扩散和空穴向 N 型区域的扩散，只剩下了非常少的来自 P 型区域的自由电子和 N 型区域空穴。这样，二极管中的反向电流就变得非常小。

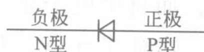


图 3-3 二极管的电路符号

图3-3 给出了二极管在电路中的符号示意图；图3-4 给出了二极管电流与电压关系的曲线图。二极管中 P 型区域侧称为“正极”，N 型区域侧称为“负极”。在大功率二极管中，正向电流可以高达 10～100A。反向电流一般很小，通常为毫微安级（10^{-9}A），甚至皮安级（10^{-12}A）。因此，二极管相当于一个电压控制的单向电流开关。当添加正向偏置电压时，允许电流正向通过；当添加反向偏置电压时，电流就变得非常小。二极管广泛应用在各种电路中。例如，用于将交流电转换成直流电的整流器、波形整流电路、峰值检波器、直流电平移动电路以及信号转换开关等等。二极管也可用于振幅调制（AM）信号的解调。

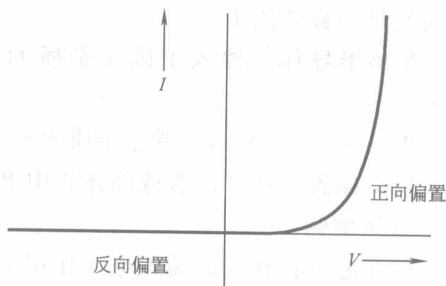


图 3-4 二极管伏安曲线图

名词解释

受主杂质：掺杂在半导体中可以产生空穴的杂质原子。例如，在硅晶体中，元素周期表中的第Ⅲ组元素就可以作为受主原子，如硼。

正极：二极管的 P 型区域侧。

负极：二极管的 N 型区域侧。

接触电势：在没有外置电压情况时，基于热平衡条件在 PN 结两端形成的内部电压。

施主杂质：掺杂在半导体中可以产生自由电子的杂质原子。例如，在硅晶体

中，元素周期表中的第V组元素就可以作为受主原子，如磷、砷、锑。

掺杂物：掺入到半导体晶体中用于调节导电性能的杂质原子。

掺杂半导体：含有用来调节导电性能的杂质原子的半导体。

非本征半导体：掺入了用来调节导电性能的杂质的半导体。

正向偏置：在二极管或晶体管中PN结两端添加的一个电压，该电压正极连接P型区域侧。

正向电流：在正向偏置电压作用下，二极管中形成的较大的电流。

空穴：在半导体两个原子的共价键中形成的电子空位。空穴是带正电的载流子，而自由电子是带负电的载流子。

本征半导体：具有一定纯度的晶体半导体，其电气特性很稳定。

多数载流子：在半导体中，占多数的载流子类型。例如，在N型半导体中，自由电子就是多数载流子。

少数载流子：在半导体中，占少数的载流子类型。例如，在N型半导体中，空穴就是少数载流子。

N型半导体：掺入了施主杂质的半导体，其自由电子数量远远大于空穴数量。

P型半导体：掺入了受主杂质的半导体，其空穴数量远远大于自由电子数量。

反向偏置：在二极管或晶体管中PN结两端添加的一个电压，该电压负极连接P型区域侧。

反向电流：在反向偏置电压作用下，二极管中形成的微小的电流。

热电压：数值上等于kT/q。其中，k为波尔兹曼常数；T为热力学温度；q为电子电量。热电压只随温度而变化，单位为V；在常温下，约为25mV。

参考文献

[1] Comer, D. J. and Comer, D. J. 2002. *Advanced Electronic Circuit Design*. John Wiley & Sons, New York, NY.

[2] Hambley, A. R. 2000. *Electronics*, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.

[3] Jaeger, R. C. and Travis, B. 2004. *Microelectronic Circuit Design with CD-ROM*. McGraw-Hill, New York.

[4] Mauro, R. 1989. *Engineering Electronics*. Prentice-Hall, Englewood Cliffs, NJ.

[5] Millman, J. and Grabelv, A. 1987. *Microelectronics*, 2nd ed. McGraw-Hill, New York.

[6] Mitchell, F. H., Jr. and Mitchell, F. H., Jr. 1992. *Introduction to Electronic Design*, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.

[7] Martin, S., Roden, M. S., Carpenter, G. L., and Wieserman, W. R. 2002. *Electronic Design*, Discovery Press, Los Angeles, CA.

[8] Neaman, D. 2001. *Electronic Analysis with CD-ROM*. McGraw-Hill, New York.

[9] Sedra, A. S. and Smith, K. C. 2003. *Microelectronics Circuits*, 5th ed. Oxford University Press, Oxford.

[10] Spence, R. and Mohammed G. 2003. *Introduction to Electronic Circuit Design*. Prentice-Hall, Englewood Cliffs, NJ.

备注

下面是一本非常好的介绍物理电子元器件的书籍：Ben G. Streetman 的《*Solid State Electronic, Devices*, 4th ed.》（Prentice-Hall, Englewood Cliffs, NJ, 1995）；另外一本介绍各种电子器件的书是 Kwork K. Ng 的《*Complete Guide to Semiconductor Devices*》（2002. Wiley-IEEE Computer Society Press, New York, NY）；介绍电路和应用的书籍是 Donald A. Neamen 的《*Electronic Circuit Analysis and Design*, 2nd ed.》（2001, Irwin, Chicago, IL）。

第4章 MOS场效应晶体管

John R. Brews

4.1 引言

金属氧化物半导体型场效应晶体管（Metal-Oxide-Semiconductor Field Effect Transistor，MOSFET）又称为“绝缘栅型场效应晶体管”，其工作原理是利用栅极上的控制电压，结合源极和漏极，依靠半导体表面的电场效应，在半导体中感应出导电沟道进行工作。栅极与其下面的半导体主体通过栅极绝缘层相互隔离，该栅极绝缘层通常为硅氧化物。半导体表面的导电沟道在半导体主体和栅极绝缘层之间形成，如图4-1所示。

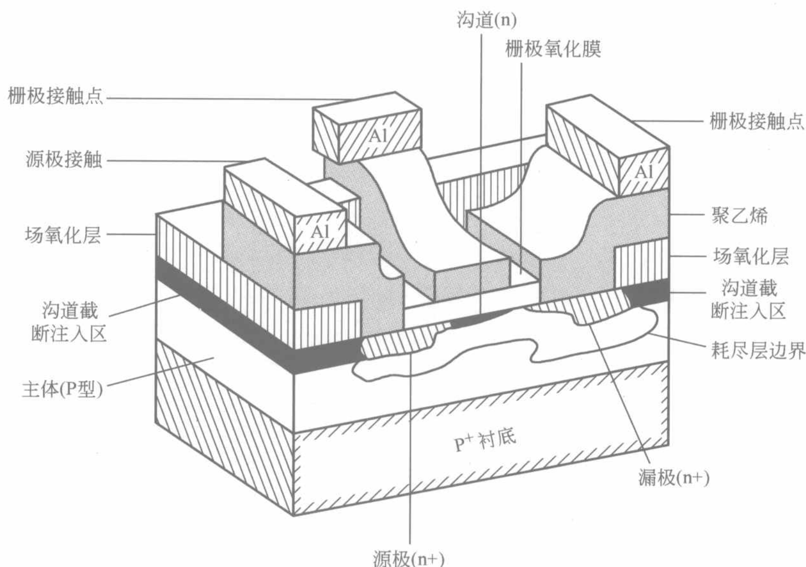


图4-1 高性能N沟道MOSFET（来源：After Brews，J. R. 1990. The submicron MOSFET. In *High-Speed Semiconductor Devices*，ed. S. M. Sze，pp. 139-210. Wiley，New York.）

如图4-1中所示，该器件通过一层厚厚的场氧化层与周围器件相互隔离，场氧化层下面是高掺杂的沟道截断注入区，沟道截断注入区用来抑制其他沟道的形成，这些沟道可能会导致该器件与周围其他器件发生连接。漏极电极处于场氧化层之上，用来降低场氧化层与主体之间的电容，该电容是寄生电容，用来降低响

应时间。该场效应晶体管的详细结构将在后面介绍。

MOSFET 的工作原理可以结合其他类型的场效应器件来理解，如结型场效应晶体管（Junction Field Effect Transistor，JFET）和金属半导体场效应晶体管（MEtal Semiconductor Field Effect Transistor，MESFET）（Hollis and Murphy，1990）。这些场效应晶体管通过控制两个欧姆接触之间横截面的方式来调节多数载流子路径上的导电能力（多数载流子是指在无电场作用的半导体中占绝对多数的载流子类型，如 N 型半导体材料中的自由电子和 P 型材料中的空穴）。这种对横截面的控制可以在沟道的任何一点进行，因此，栅极可以处于任何位置，而且无需扩展到整个沟道。

与上面这些场效应晶体管类似的是掩埋沟道型、耗尽型或常开型 MOSFET，这类场效应晶体管包含一个掺杂物表面层，该表面层中的掺杂物类型与源极和漏极中的相同（与场效应晶体管器件主体类型相反）。因此，这类场效应晶体管在源极和漏极之间拥有一个内置的或常开的导电沟道，当栅极耗尽多数载流子时，该沟道会缩小。

而恰恰相反，真实的 MOSFET 是一种增强型或常闭型器件。这种类型的场效应晶体管是常闭型的，因为器件主体与源极和漏极之间都形成了 PN 结，这样在它们之间无法形成多数载流子电流，取而代之的是由少数载流子形成的电流。在稍后的讨论中可以发现，由于栅极偏置电压的作用（高于最低门限值：开启电压），少数载流子会被牵引到了表面的导电沟道中，在源极和漏极之间形成了一个导电路径。这样，在栅极和导电沟道之间就形成了一个由栅极绝缘层隔离的电容。当栅极这一侧堆积电荷时，在导电沟道一侧也会从源极和漏极之间吸引等量的异性电荷，以达到平衡。栅极上堆积的电荷越多，形成的导电沟道就越多，导电能力也就越强。由于栅极“导致”了沟道的产生，为确保导电的持续性，栅极必须扩展到源极和漏极之间的整个范围区域。

MOSFET 中的导电沟道是由于栅极的吸引以及沟道和栅极之间的绝缘层而形成的，该绝缘层用来阻止少数载流子向栅极渗漏。因此，MOSFET 器件只能由可以提供良好绝缘层的材料系统制造，而且最著名的是硅-二氧化硅混合物。这个必须提供良好绝缘层的要求对于 JFET 和 MESFET 来说并不重要，因为在 JFET 和 MESFET 中，栅极的作用是用来排斥多数载流子的，而不是用来吸引少数载流子的。因此，在 GaAs 系统中，由于良好的绝缘层和其他的元器件或制造要求不一致，我们通常使用 MESFET。

最新开发的 GaAs 系统采用的是异质结构场效应晶体管（Heterostructure Field Effect Transistor，HFET）（Pearton and Shah，1990），该类型场效应晶体管由不同含量的 Al、Ga 和 Si 或者 In、Ga、P 和 As 构成。这类场效应晶体管通常通过分子束外延技术或有机金属气相外延技术制造。HFET 具有很多种结构，其

中最著名的是调节掺杂场效应晶体管（MOdulation Doped Field Effect Transistor, MODFET)。HFET是场效应晶体管，但不是MOSFET，因为它的栅极只是简单用来调节欧姆接触之间已经预先存在的导电沟道中的载流子密度。这种导电沟道与栅极绝缘层无关，它是自然形成的，用来调节各层之间的平衡，就像一个PN结中形成的耗尽层。最后形成的沟道非常靠近栅极，并产生类似MOSFET中的栅极控制作用。

硅基MOSFET一出现就受到了极大的欢迎，原因在于硅-二氧化硅系统提供了一个具有低缺陷密度而且性能稳定的界面；另外，氧化物对于环境中的污染具有阻隔渗透的作用，并具有很强的抗损硬度，而且容易均匀生长或再生（Nicollian and Brews，1992)。上述这些特征可以通过平版印刷技术来轻松实现，而且还可以生产出具有元器件小型化、大规模、高可靠性和低成本等特点的集成电路(IC)。由于MOSFET和高密度制造工艺密切相关，因此本章的一个重点就是讨论元器件的小型化问题。

MOSFET的另外一个优势是既可以利用电子也可以利用空穴作为导电沟道的载流子，形成两种类型的MOSFET。在所谓的互补MOS（Complementary Metal Oxide Semiconductor Transistor，CMOS）技术中使用这两种MOSFET时，如果电流路径中至少包含了一个串联电路（该串联电路由这两种MOSFET构成)，那么这个电路就不能连接直流电源，因为在稳定状态下直流电源会导致只有一种载流子存在，即两种载流子不能同时存在。当然，在调试电路的过程中，转换器件时需要切断电源。这种在N沟道或P沟道器件选择上的灵活性使得大规模电路的制造可以处于一个低功耗的水平。因此，复杂系统在制造时可以避免昂贵的封装和冷却要求。

4.2 伏安特性

本小节介绍了MOSFET伏安特性的来源（Annaratone，1986；Brews，1981；and Pierret，1990)，并给出了定性的讨论。

1. 强反型特性

图4-2给出了源极到漏极的电流 I_D 与加载在源极与漏极之间电压 V_D 的关系(MOSFET的 *I-V* 曲线图)。当电压 V_D 值较低时，电流 I_D 随着电压 V_D 的增加呈近似线性变化关系，其电气特性近似于一个阻值受栅极电压 V_G 控制的简单电阻：当栅极电压升高并对沟道载流子产生更强的吸引时，导电沟道就会变得更宽，导电沟道中的载流子也会也来越多，其电阻率 R_{ch} 也会下降。因此，V_D 值较低时，栅极电压 V_G 越大，导电沟道中的电流就越强。

当电压 V_D 值较高时，*I-V* 曲线就会趋向平缓，电流也会随着漏极电压的增

加而基本保持不变。此时，MOSFET 处于饱和状态。产生饱和状态的原因很多，其主要原因就是加载在导电沟道上的漏极电压。如果源极与漏极之间的距离很短，或者相邻，或者只相差 1μm，那么一般的漏极电压就可以在导电沟道产生高于 10^4V/cm 的场强。这样，载流子就具有了足够的能量并通过晶体中硅原子的振动来释放能量（光子发射）。到最后，载流子的速度随着电压场强增加时就不会出现很大的变化，硅 MOSFET 中的饱和速度大约为 $v_{sat} \approx 10^7$cm/s。由于载流子速度在饱和状态下不会再随着 V_D 继续增加，因此，电流也就达到了饱和状态。

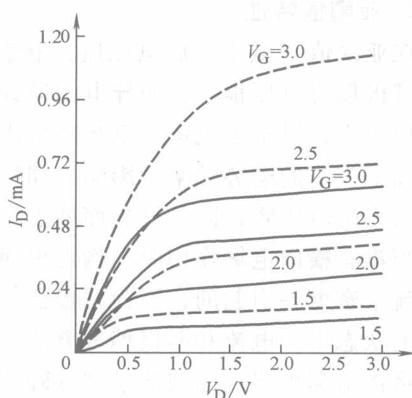


图 4-2　各种栅极偏置 V_G 下漏极电流-漏极电压（I_D-V_D）关系曲线

注：虚线是指长沟道器件，其饱和状态下的电流随栅极偏置呈二次方增加关系；实线是指短沟道器件，其状态接近速度饱和，而且饱和电流随栅极偏置呈线性增加关系，在本小节中将进行讨论

对于长沟道器件来说，其 *I-V* 曲线出现饱和状态的原因会有所不同。我们将绝缘层沟道界面的电势称为“表面电势”。无论沟道的源极终端处器件表面电势是多少，该电势从源极终端到漏极终端将会逐渐变大，直至达到 V_D，因为漏极电势为 V_D，高于源极电势。另一方面，由于栅极上的各处电势相同，因此，栅极与源极之间的电势差比栅极与漏极之间的电势差要大。相比来说，源极氧化层的面积比漏极大，因此，漏极上能负载的电荷就少。这样一来，栅极上对电荷吸引能力也变小，从而导致漏极终端沟道中的载流子数目变少，最后导电沟道的电阻率就增加了。简单来说，$I_D \approx V_D / R_{ch}$；其中，电阻率 $R_{ch} = R_{ch}(V_D)$ 随电压 V_D 的增加而增加。因此，*I-V* 曲线在经过了初始的直线变化关系后，就会变平缓并达到饱和状态。

对于长沟道器件和短沟道器件来说，*I-V* 曲线上的另一个区别是对栅极电压的依赖程度。对于长沟道器件而言，饱和电流值 $I_{D,sat}$ 会随着栅极偏置电压的增加而呈平方增加关系，这是因为导电沟道中的载流子数目和 $V_G - V_{TH}$（其中，V_{TH} 为开启电压）呈比例变化关系，这一点将在后面讨论。而沟道的电阻率 $R_{ch} \propto 1/(V_G - V_{TH})$，且漏极偏置电压达到饱和时的值为 V_G。因此，$I_{D,sat} = V_D / R_{ch} \propto 1/(V_G - V_{TH})^2$。这样，我们就得到了一个呈二次方变化的关系。但是，当载流子速度达到饱和时，这种电流和漏极偏置电压之间的相互关系就会得到抑制。因为载流子速度在饱和状态时被固定在了 v_{sat}，而 $I_{D,sat} \propto v_{sat}/R_{ch} \propto (V_G - V_{TH})v_{sat}$，从而 $I_{D,sat}$ 就和栅漏偏置电压之间成了线性变化关系。因此，对于短沟道器件来说，其产生的电流就没有长沟道器件大。

2. 亚阈值特性

在亚阈值（低于开启电压值）状态下，我们可以发现伏安特性有很大不同，因为栅极偏置电压很小，而导电沟道处于弱反型状态。此时，导电沟道中的载流子数目很少，以至于无法对电势产生影响，而且这些载流子还必须适应由电极和掺杂离子产生的电势带来的影响。同样在亚阈值状态下，任何微小电流在界面上都无法产生电压降，因此该界面处于等势状态。

当没有横向电场作用时，沟道里面的载流子只会做简单的扩散运动，这是由于载流子密度差引起的，因为漏极会导致载流子密度在沟道的漏极终端减小。在亚阈值状态下，电流和漏极偏置电压无关。但是一旦偏置电压超过数十毫伏，就足够将沟道漏极终端的载流子密度降到接近零。

但是，对于短沟道器件来说，由于源极和漏极靠得太近以至于它们开始共享对栅极电势的控制。如果这种控制作用太强，在漏极电压和亚阈值之间就会出现相互依赖关系，这是我们不希望发生的，因为它会增加 MOSFET 的空闲电流，并会导致漏极电压和开启电压之间产生相互依赖关系。

对于理想设计的器件来说，尽管漏极电压和亚阈值之间不存在相互关系，但是栅极偏置电压和亚阈值之间呈指数变化关系。相比半导体主体而言，表面能量由于栅极的作用而被降低了。如果表面电势 ϕ_S 低于主体电势，载流子密度就会在波尔兹曼因子 $\exp(q\phi_S/kT)$ 的作用下上升，该因子与主体浓度有关。其中，kT/q 为热电压，在 290K 时，$kT/q \approx 25\text{mV}$。由于 ϕ_S 和 V_G 约呈比例关系，因此载流子密度和亚阈值状态下的电流值跟 ϕ_S 之间的指数关系就直接可以转换成与 V_G 之间的指数关系。

4.3 重要器件参数

很多 MOSFET 参数对于 MOSFET 的性能非常重要，本节将特地从数字 IC 的角度来讨论其中的部分参数。

1. 开启电压

开启电压是指导电沟道刚刚形成时的栅极电压值 V_{TH}，处于该电压下的器件开始从关闭状态转换到导通状态，而电路也会产生一个电压摆动，该摆动的电压值范围覆盖了开启电压值。这样，开启电压就可以用来作为电路操作的开关，还可以在器件处于关闭状态时用来监测是否漏电或断电。

接下来将给出开启电压的准确定义，并详细介绍器件的内部掺杂情况以及其他器件参数，如氧化层厚度和平带电压。

开启电压由氧化层厚度 d 和主体掺杂度控制。为了控制主体掺杂度，常常使用离子注入方法，以便掺杂离子密度不是简单的均匀分度（即单位体积内的离

子数为 N_B），而是在均匀分布的基础上再叠加注入离子密度。为了评估开启电压，我们必须了解当栅极偏置电压从闭合状态变化到临界状态时在栅极下面的半导体中所发生的变化。

我们可以想象出栅极偏置电压从闭合状态变化到临界状态时，首先产生的结果是抑制了多数载流子，并形成一个表面耗尽层，如图 4-1 所示。在耗尽层中，几乎不存在载流子，但是存在掺杂离子。在 N 型材料中，这些掺杂离子是带正电的施主杂质，并且在电场作用下无法移动，因为它们被固定在硅晶格当中，并取代了硅原子的位置。而在 P 型材料中，这些掺杂离子是带负电的受主杂质。因此，任何添加到栅极并使得栅极电压接近开启电压的电荷都会导致耗尽层宽度的增加，增加的宽度足以通过硅晶体耗尽层中等量的但电荷相反的掺杂离子来平衡栅极电荷。

扩展的耗尽层持续平衡栅极上多余的电荷，直到达到开启电压，然后用来平衡的电荷就会产生变化：在超过开启电压值之后，任何添加的栅极电荷就会被不断增加的强反型层或沟道平衡。当满足以下条件时，耗尽层和相应的反型层之间的界面就会出现临界状态：

$$\frac{dqN_{inv}}{d\phi_S}=\frac{dQ_D}{d\phi_S} \tag{4-1}$$

式中，$d\phi_S$为栅极电荷增加时表面电势产生的微小变化；qN_{inv}为单位面积上的反型层电荷；Q_D为单位面积上的耗尽层电荷。

根据式（4-1），在临界状态下，两种响应是相等的。更具体地说，qN_{inv}的增加是呈指数变化；也就是说，电荷的增加率和 qN_{inv}是呈比例关系。因此，当 qN_{inv}增加时，式（4-1）左边也会随之增加。从另一个方面来说，Q_D和 ϕ_S是平方根关系，这意味着 ϕ_S的增加速度没有 Q_D快。因此，在临界状态下，当表面电势增加时，式（4-1）左边也会随着 qN_{inv}增加，直到满足式（4-1）。当超过开启电压后，qN_{inv}将随 ϕ_S的增加而呈指数增加，Q_D忽略不计，此时 qN_{inv}的变化占主要地位。类似的，在低于开启电压时，qN_{inv}将随 ϕ_S的减小而呈指数减小，qN_{inv}可以忽略不计，此时 Q_D的变化占主要地位。这种变化的突然性就是我们在描述 MOSFET 转换时所说的专用术语"临界状态"。

在利用式（4-1）来推断开启电压规律时，我们还必须知道 qN_{inv}和 Q_D的表达式。假设界面的能量比栅极上电荷的能量低，界面的少数载流子密度就比半导体主体上大，甚至低于临界状态。在临界状态以下以及接近临界状态时，根据式（4-1），在 N 沟道器件中单位面积沟道中的电荷数 N_{inv}可以由下式（Brews，1981）给出：

$$N_{\mathrm{inv}} \approx d_{\mathrm{INV}} \frac{n_i^2}{N_{\mathrm{B}}} \exp^{q(\phi_{\mathrm{S}} - V_{\mathrm{S}})/kT} \tag{4-2}$$

式中，n_i是指单位体积本征载流子的密度，在290K时，硅晶体中$n_i \approx 10^{10}/\mathrm{cm}^3$；$V_{\mathrm{S}}$是指主体反偏电压；$d_{\mathrm{INV}}$是指界面中少数载流子离表面的有效深度，由下式定义：

$$d_{\mathrm{INV}} = \frac{\varepsilon_{\mathrm{s}} kT/q}{Q_{\mathrm{D}}} \tag{4-3}$$

式中，Q_{D}是指由于带电掺杂离子的缘故而在耗尽层单位面积上产生的电荷，掺杂区域不存在任何载流子；ε_{s}是指半导体的介电常数。

式（4-2）给出了单位面积内少数载流子的纯密度，其值等于单位体积内少数载流子的密度$\frac{n_i^2}{N_{\mathrm{B}}}$与少数载流子分布深度$d_{\mathrm{INV}}$和常见波尔兹曼因子$\exp(q(\phi_{\mathrm{S}} - V_{\mathrm{S}})/kT)$的乘积。该因子说明了半导体主体之上单位体积内低能量界面密度增加的程度。深度d_{INV}与界面附近的载流子分布有关，其值近似（弱饱和状态时有效）为远离氧化膜的少数载流子的衰退指数。在这个近似值中，d_{INV}是少数载流子密度的重心。举例来说，在290K时，均匀掺杂密度为10^6离子/cm^3的场效应晶体管，结合式（4-7）可以得出临界状态时的表面电势（$\phi_{\mathrm{TH}} = 0.69\mathrm{V}$）；同时利用式（4-2）可以计算出临界状态时在耗尽层的电荷密度$Q_{\mathrm{D}}/q = 3 \times 10^{11}$电荷/$\mathrm{cm}^2$。与$Q_{\mathrm{D}}$对应的$d_{\mathrm{INV}} = 5.4\mathrm{nm}$，临界状态时的载流子密度为$N_{\mathrm{inv}} = 5.4 \times 10^9$电荷/$\mathrm{cm}^2$。

下一步我们将利用开启电压的定义来计算单位面积上的耗尽层电荷Q_{D}。对于注入离子的情况来说，Q_{D}由两部分组成（Brews，1981），如下式所示：

$$Q_{\mathrm{D}} = qN_{\mathrm{B}}L_{\mathrm{B}}(2(q\phi_{\mathrm{TH}}/kT - m_1 - 1))^{\frac{1}{2}} + qD_I \tag{4-4}$$

式中，第一部分是Q_{B}，Q_{B}是指耗尽层中来自主体掺杂原子的耗尽层电荷，其宽度在第一次离子注入时变小，也就是m_1由注入重心x_{C}给出，如下式所示：

$$m_1 = \frac{D_I x_{\mathrm{C}}}{N_{\mathrm{B}} L_{\mathrm{B}}^2} \tag{4-5}$$

第二部分是附加电荷，附加电荷是由于耗尽层中的注入离子密度引起的，数值为单位面积的D_I。德拜长度L_{B}的定义如下：

$$L_{\mathrm{B}}^2 \equiv \left[\frac{kT}{q}\right]\left[\frac{\varepsilon_{\mathrm{s}}}{qN_{\mathrm{B}}}\right] \tag{4-6}$$

式中，ε_s是指半导体的介电常数。

德拜长度是指当 $D_I=0$ 和耗尽层宽度为0时，表面电势在半导体主体中的渗透程度。

式（4-2）给出了 qN_{inv}，公式（4-4）给出了 Q_D，式（4-7）则定义了临界状态时的表面电势 ϕ_{TH}：

$$\phi_{TH}=2(kT/q)\ln(N_B/n_i)+(kT/q)\ln\left[1+\frac{qD_I}{Q_B}\right] \tag{4-7}$$

式中，Q_B是指耗尽层中主体掺杂为 N_B时的单位面积耗尽层电荷；qD_I是指在耗尽层边缘和反型层边缘之间注入离子时的单位面积耗尽层电荷。

由于 ϕ_S在 ϕ_{TH}基础上的微小增加会直接导致 qN_{inv}迅速上升，从而可以对栅极电荷或栅极电压产生很强的平衡作用，因此 ϕ_S没有 V_G-V_{TH}增加得快。但是，在强反型层中，$N_{inv}\approx10^{12}$ 电荷/cm^2，因此在强反型层中 ϕ_S 比 ϕ_{TH} 大，约为 $7kT/q$。

式（4-7）说明了出现临界状态时均匀掺杂的情况（无离子注入，$D_I=0$），此时 $\phi_S=\phi_{TH}=2(kT/q)\ln(N_B/n_i)\equiv2\phi_B$。但是对于非均匀掺杂的情况来说，就需要很高的表面电势；因为假设正常的离子注入时，D_I 为正数，表示需要提高掺杂浓度。离子注入会导致所需表面电势增加，这是因为表面的电场很大，导致反型层变窄，沟道长度变短 $\phi_S=2\phi_B$。因此，增加相应的 qN_{inv} 时，需要更高的表面电势，以满足式（4-1）。如果足够比例的注入离子被限制在反型层自身中，那么式（4-1）就不适用了。但是，在现实情况中，注入离子不可能被限制在和反型层厚度（数十纳米）相当的距离内，因此，式（4-7）适用于所有现实情况。

当已知表面电势 ϕ_{TH}时，如果我们知道了氧化层场强 F_{ox}，通过在半导体主体和氧化层之间添加电势差，就可以得到临界状态时的栅极电压，式为 $\Phi_{TH}=\phi_{TH}+F_{ox}d$，这里 d 是指氧化层厚度，F_{ox} 由下面的高斯定律给出：

$$\varepsilon_{ox}F_{ox}=Q_D \tag{4-8}$$

在研究开启电压时，我们会遇到两种复杂情况。第一种情况是，在临界状态下，由于半导体主体和栅极材料之间工作原理的差异，栅极电势 Φ_{TH}和栅极电压 V_{TH}会有所不同。这种差异会导致当MOSFET应用在电路中并开始产生电荷转移时，两种材料之间会很自然地产生电荷交换。因此，在给元器件添加任何电压之前，在栅极和主体之间都会由于自然的电荷转移而存在电势差。第二种影响开启电压的复杂情况是在绝缘体和绝缘体-半导体界面中有电荷存在。以上这些对整体电荷平衡产生影响的非理想情况是由于晶体陷阱和器件运行过程中的固定电荷

合并在一起而导致的。

通常，界面陷阱电荷是可以忽略不计的（在硅 MOSFET 中 $<10^{10}/\mathrm{cm}^2$），而其他对开启电压产生的非理想影响可以由“平带电压 V_{FB}”来说明，平带电压 V_{FB}用来校正这些影响产生的栅极偏置。然后，根据式（4-8）加上 $F_{ox}=(V_{TH}-V_{FB}-\phi_{TH})/d$，我们可以得到：

$$V_{TH}=V_{FB}+\phi_{TH}+Q_D\frac{d}{\varepsilon_{ox}} \tag{4-9}$$

式（4-9）给出了非均匀掺杂情况下的 V_{TH}值；其中，ϕ_{TH}可以由式（4-7）得出；临界状态下的 Q_D 可以由式（4-4）得出。如果单位面积上界面陷阱电荷值是负数，那么专门用来表示单位面积上界面陷阱电荷的 Q_{IT}必须添加到式（4-9）中的 Q_D 上。

从式（4-4）~式（4-7），我们可以发现，开启电压只与注入掺杂离子的两个分布参数相关：在反型层和耗尽层边缘之间的区域通过注入离子产生的净电荷 qD_I 和由注入产生的这种电荷分布重心 x_C。因此，很多注入方式可以产生相同的开启电压，这些注入方式包括以 x_C 为重心的 δ 函数极限值的穿刺注入（单位面积上注入数量为 D_I）以及相同数量和重心的箱形矩形分布，即一个宽度 $x_W=2x_C$ 和体积密度为 D_I/x_W 的矩形分布（当然，这里的 x_W 不能超过临界状态时耗尽层的宽度，以确保等式有效；x_C 不能处于反型层宽度范围中）。这种对细节很弱的依赖性使得器件在满足其他要求时具有了很大的灵活性，例如对截断电流的控制。

如前所述，如果栅极偏置 $V_G>V_{TH}$，任何高于开启电压值的栅极电压电荷大部分都会被反型层电荷平衡掉。因此，附加的氧化层场强为 $(V_G-V_{TH})/d$，根据高斯定律，该场强和反型层载流子密度相关，其关系为

$$\varepsilon_{ox}(V_G-V_{TH})/d\approx qN_{inv} \tag{4-10}$$

式（4-10）说明了，器件开启后，沟道长度和 V_G-V_{TH}成正比关系，V_G-V_{TH}是本章中常见的近似值。因此，当平衡电荷从耗尽层转移到反型层时，N_{inv}与栅极电压之间的关系也从亚阈值时的指数关系变成了开启电压之上的线性关系。

根据电路分析原理，式（4-10）为 V_{TH}给出了一个很实用的定义，因为它非常适合伏安特性曲线。如果用这个定义替代式（4-1）中平衡电荷的定义，那么式（4-1）和式（4-7）将可以近似为 ϕ_{TH}。

2. 电荷驱动能力和 $I_{D,sat}$

MOSFET 的电荷驱动能力与其在给定栅极偏置电压下能提供的电流成正比。因此我们可以推断，电流越大，电路运行越快。这里，由电流可以得到支配

MOSFET 电路的响应时间。

MOSFET 电流取决于沟道中载流子的密度或 $V_G - V_{TH}$，如式（4-10）所示。对于长沟道元器件，电荷驱动能力还取决于沟道长度。沟道长度 L 越短，电荷驱动能力越强，因为沟道电阻率与沟道长度成正比。虽然这个原理过于简单，但是我们可以假设 MOSFET 在驱动电路负载时处于饱和状态，这样一来，我们在讨论如何使 MOSFET 运行更快时，可以有一个更清晰的思路，从而避免复杂的数学运算。假设在大部分转换阶段，MOSFET 都处于饱和状态，电荷驱动能力就和饱和电流成正比，如下式所示：

$$I_{D,sat} = \frac{\varepsilon_{ox} Z \mu}{2dL}(V_G - V_{TH})^2 \tag{4-11}$$

式中，饱和电流 $I_{D,sat}$ 是高漏极偏置电压下 I-V 曲线产生饱和特性的主要因素；Z 是指正常电流方向沟道的长度。从上面公式可以看出，对于长沟道元器件来说，电荷驱动能力和 $V_G - V_{TH}$ 成二次方关系，和 d 成反比关系。

式（4-11）的结果适用于长沟道元器件。而对于短沟道元器件（见图 4-2），由漏极产生的更大场强导致了“饱和速率”的出现，因此，$I_{D,sat}$ 可以近似为

$$I_{D,sat} \approx \frac{\varepsilon_{ox} Z v_{sat}}{d} \frac{(V_G - V_{TH})^2}{V_G - V_{TH} + F_{sat} L} \tag{4-12}$$

式中，v_{sat} 是指载流子饱和速度，在硅晶体中，在 290K 时约为 10^7cm/s；F_{sat} 是指饱和速度下的场强，在硅 MOSFET 中，电子在饱和速度下的场强约为 5×10^4 V/cm，空穴在饱和速度下的场强约 $\geq 10^5$ V/cm。

为了使式（4-12）和式（4-11）在 L 上取得一致，在硅 MOSFET 中，我们需要电子的 $\mu \approx 2v_{sat}/F_{sat} \approx 400\text{cm}^2/(\text{Vs})$，这个值只是一个近似值。但是我们可以发现，在亚微型的沟道长度定义中，$I_{D,sat}$ 逐渐与沟道长度 L 无关，而和 $V_G - V_{TH}$ 之间逐渐变成线性关系而非平方关系，如式（4-23）所示。式（4-23）说明了，当 $(V_G - V_{TH})/L \geq F_{sat}$ 时，速率就可以达到足够饱和的状态，例如当 $V_G - V_{TH} = 2.3$V，$L \leq 0.5\mu$m 时。

为了说明 $I_{D,sat}$ 和栅极响应时间 τ_G 之间的关系，我们假设一个 MOSFET 驱动另一个相同的 MOSFET 并将其作为负载电容。那么在栅极响应时间 τ_G 内，由式（4-12）产生的电流将对这个负载电容进行充电，并产生一个电压 V_G，τ_G 由下式给出（Shoji，1988）：

$$\begin{aligned}\tau_G &= C_G V_G / I_{D,sat} \\ &= \left[\frac{L}{v_{sat}}\right]\left[1 + \frac{C_{par}}{C_{ox}}\right]\left[\frac{V_G(V_G - V_{TH} + F_{sat} L)}{(V_G - V_{TH})^2}\right]\end{aligned} \tag{4-13}$$

式中，C_G 是指 MOSFET 栅极的电容值，$C_G = C_{ox} + C_{par}$；C_{ox}为 MOSFET 氧化层电容，$C_{ox} = \varepsilon_{ox} ZL/d$；$C_{par}$为栅极电容的寄生电容（Chen，1990）。

寄生电容 C_{par}主要是由于栅极与源极和漏极存在重叠而产生的，另外还与边缘效应场强和沟道边缘电容有关。当沟道长度 L 较短时，寄生电容 C_{par}是 C_G 的主要部分。而当沟道长度 L 减小时，栅-漏调整技术就可以控制 C_{par}。典型的例子是，当 $V_{TH} \approx V_G/4$ 时，

$$\tau_G = \left[\frac{L}{v_{sat}}\right]\left[1 + \frac{C_{par}}{C_{ox}}\right]\left[1.3 + 1.8\frac{F_{sat}L}{V_G}\right] \tag{4-14}$$

因此，在本征能级时，栅极响应时间是电子从源极转移到漏极时间的数倍，当速度饱和时，电子从源极转移到漏极的时间为 L/v_{sat}。当沟道长度 L 较短时，栅极响应延迟时间随 L 线性减小；当沟道长度 L 较长时，栅极响应时间随 L 呈二次方增加，同时它还与 V_G 随 L 变化的速度有关。

栅极响应时间不是元器件转换时的惟一时间延迟，因为漏极-主体的 PN 结在 MOSFET 的状态转换过程中也会进行充电或放电（Shoji，1988）。因此，我们还必须考虑漏极响应时间 τ_D。根据式（4-13），我们假设漏极电容 C_D 由饱和状态下 MOSFET 提供的电压进行充电，因此

$$\tau_D = C_D V_G / I_{D,sat} = \left[\frac{C_D}{C_G}\right]\tau_G \tag{4-15}$$

式（4-15）说明了，当 L 减小时，τ_D 也会随着 τ_G 的增加而增加，而 C_D/C_G 不会随着 L 的减小而增加。但是，$C_{ox} \propto L/d$ 和 C_{par}的主要部分即重叠分布电容使得 $C_{par} \propto L_{ovlp}/d$；这里，$L_{ovlp}$大约是栅极在源极和漏极上重叠长度的三倍（Chen，1990）。这样，$C_G \propto (L + L_{ovlp})/d$，为了使 C_D/C_G 不会随着 L 的减小而增加，C_D 或氧化层厚度 d 必须随 L 的减小而减小。

先进的电路设计会选择减小 C_D。例如，各种凸起的漏极电路设计通过利用较厚的氧化层将大部分漏极区域与主体进行隔离来减小漏极-主体电容。由紧邻沟道区域的耗尽层边缘宽度产生的漏极电容更加难以处理，因为耗尽层边缘在器件微型化过程中会被特意减小以避免产生“短沟道效应”；也就是说，漏极会在栅极电压控制的作用下对沟道产生影响。最后导致的结果是，耗尽层边缘产生的漏极电容会随着器件的微型化而逐渐增加，除非 PN 结深度减小。

式（4-14）和式（4-15）说明了当沟道长度 L 减小时，器件响应时间会减小。减小氧化层厚度不会使 τ_G 增加，但是式（4-15）给出了 τ_D 增加的可能性。环形电路振荡器是一个包含奇数个变换器的闭环电路，该闭环电路是一个测试电路，其性能主要与 τ_D 和 τ_G 相关。例如，当沟道长度为 0.1μm 时，环形电路振荡器每段的栅极延迟约为 12ps/每段；当沟道长度为 0.5μm 时，栅极延迟约为

60ps/每段。

在电路中，互连电容和“输出端数”（各种 MOSFET 负载）会使响应时间增加并超过器件的响应时间，即使将寄生电容也算在内。因此，我们必须考虑互连延迟 τ_{INT}。尽管集总模型和式（4-15）说明了 $\tau_{INT} \approx (C_{INT}/C_G)\tau_G$，但是互连长度要求对应一个分布式的模型。因此，互连延迟为

$$\tau_{INT} = R_{INT}C_{INT}/2 + R_{INT}C_G + (1 + C_{INT}/C_G)\tau_G \tag{4-16}$$

式中，R_{INT}是指互连电阻；C_{INT}是指互连电容。

与此同时，我们还假设互连电路中还包含一个处于饱和状态的 MOSFET 作为驱动器，另一个 MOSFET 作为负载 C_G。当 R_{INT}较小时，由式（4-16）可以看出，此时 τ_{INT}由公式中的最后一项决定，类似于式（4-13）和式（4-15）。但是与式（4-15）中的 C_D/C_G 比例关系不同的是，在式（4-16）中，当 L 减小时，C_{INT}/C_G 很难减小甚至是维持原值不变。值得注意的是，$C_G \propto Z(L + L_{ovlp})/d$。因此，当 L 减小时，C_{INT}/C_G 会增加，这是因为如果 R_{INT}没有非常不切实际地增加，互连横截面就不会减小。更加糟糕的是，随着 L 的不断减小，芯片的尺寸通常会不断增加，同时还会导致线路长度不断增加，而且即使是互连横截面宽度不变，固定交叠部分的 R_{INT}也会增加。因此，当 L 减小时，互连延迟就成了最大的问题了。最显而易见的控制 C_{INT}/C_G 的方法就是增加元器件宽度 Z，这样 $C_G \propto Z(L + L_{ovlp})/d$ 在 L 减小时就可以保持不变。更好的方法是通过层叠元器件来增加宽度 Z（Chen，1990；Shoji，1988）。但是，对于任何一种方法来说都需要额外的区域来减小封装密度，而减小封装密度是在减小 L 时的首要目标。另外一个可选方案就是减小氧化层厚度，而减小氧化层厚度是一个技术目标。

3. 跨导

另外一个重要的器件参数是小信号跨导 g_m（Malik，1995；Sedra and Smith，1991；Haznedar，1991）；跨导决定了漏极输出电流，而漏极输出电流随给定栅极输入电流而变化，也就是小信号增益：

$$g_m = \left.\frac{\partial I_D}{\partial V_G}\right|_{V_D = \text{const}} \tag{4-17}$$

基于差异的链式法则，饱和状态下的跨导与小信号跃迁或单位增益频率相关，单位增益频率决定了小信号电流增益$|l_{out}/l_{in}| = g_m/(\omega C_G)$ 在多高的频率 ω 下降到统一值。根据链式法则，g_m如下所示：

$$g_m = \frac{\partial I_{D,sat}}{\partial Q_G}\frac{\partial Q_G}{\partial V_G} = \omega_T C_G \tag{4-18}$$

式中，C_G 是指器件的氧化层电容，$C_G = \partial Q_G/\partial V_G|_{V_D}$，其中 Q_G 是指栅极电荷。

频率 ω_T是小信号、高频元器件在忽略了寄生电阻后的衡量标准。结合式（4-12），在式（4-18）中我们可以发现跃迁频率同样和式（4-14）中的转移时间 L/v_{sat}相关，因此，数字电路和小信号电路的速度都和该参数相关。

4. 输出电阻和漏极电导

在小信号电路中，MOSFET 的输出阻抗 r_o（Malik，1995；Sedra and Smith，1991）对于限制放大增益非常重要。该输出阻抗与饱和状态下的小信号漏极电导率相关，公式如下：

$$r_o = \frac{1}{g_D} = \frac{\partial V_D}{\partial I_{D,sat}}\bigg|_{V_G = const} \tag{4-19}$$

如果 MOSFET 单独作为放大器，其等效负载为 R_L，那么增益就变成了：

$$\left|\frac{v_o}{v_{in}}\right| = g_m \frac{R_L}{R_L + r_o} \leqslant g_m R_L \tag{4-20}$$

式（4-20）描述了当 r_o 逐渐减小至 R_L 时，增益是如何减小的。

随着元器件的微型化，r_o 会逐渐减小，而增益会逐渐提高，这是很多因素共同作用的结果；其中，主要因素就是在适当的漏极偏置电压下对沟道长度的调整。当沟道长度减小时漏极电压会增加，同时漏极周围的耗尽层区域就会向源极扩展，导致沟道长度 L 和漏极偏置电压相关。第二个因素就是在较大的漏极偏置电压下对反型层电荷密度的控制。在短沟道元器件中，这种控制作用会和栅极控制产生竞争。这和稍后讨论的亚阈性能原理相同。在更大的漏极偏置电压下，载流子的增加会使 r_o 进一步减小。

在数字转换器中，较小的 r_o 会加剧电压摆动幅度，该电压摆动用来在输出电压中产生跃迁。这种对电压摆动的加剧会增加功率损耗，因为在跃迁过程中，电流会产生穿刺并降低噪声容限（Annaratone，1986）。但是，这还不是数字电路中元器件微型化首要关心的问题。因为小信号电路比数字电路对 r_o 更敏感，因此小信号电路中的 MOSFET 无法制造得像数字电路中那样小。

4.4　元器件微型化的局限性

MOSFET 成功的主要原因是其在很小的尺寸下对各种应用具有兼容性。目前，0.5μm 的沟道长度（源极至漏极的距离）已经可以实现，更小的 0.1μm 沟道已经在有限数量的测试电路中实现，如环形电路振荡器。在本节中，我们将介绍在元器件微型化过程中必须考虑的局限性（Brews，1990）。

1. 亚阈控制

当 MOSFET 处于截断状态时，也就是说，MOSFET 处于亚阈状态时，随电源

电压变化的截断电流不能太大，以免产生功率消耗以及表面隔离节点的放电（Shoji，1988）。但是，在小元器件中源极和漏极相隔很近，这样在源极和漏极之间就会存在一种隐患，即源极和漏极无须通过栅极和沟道而直接相互影响。在极端情况下，漏极甚至可能直接将源极电流牵引过来，即使栅极处于截断状态（即穿通）。另外，仅次于极端但也不希望发生的情况就是，漏极和栅极联合控制了沟道中载流子的密度（漏极引起的势能下降或漏极对开启电压的控制）。在这种情况下，MOSFET 的截断或开启就无法由栅极来单独控制了，其发生转换的栅极电源范围就变得很广了，该范围和漏极电压有关。在上面这种情况下，电路的可靠性设计是非常复杂的，而且不能对设计误差进行测试。因此，在设计 MOSFET 时，必须确保漏极偏置与亚阈性能无关。

衡量源极和漏极影响的标准之一是耗尽层中相关 PN 结的宽度。这种 PN 结的耗尽层是指所有载流子都被耗尽或者由于 PN 结的电压而导致载流子都被驱赶的区域。PN 结的电压包括 PN 结两端的偏置电压以及当 P 型区域和 N 型区域接触时自然产生的内部电势差。突变结的耗尽层宽度 W 与电势差 V 以及单位体积内的掺杂离子浓度 N 有关，公式为

$$W=\left[\frac{2\varepsilon_{s}V}{qN}\right]^{\frac{1}{2}} \tag{4-21}$$

为了避免出现亚阈问题，常见的处理规则是确保沟道长度大于最小长度，最小长度 $L_{\min}$ 与结的深度 r_j 和氧化层厚度 d 以及源极和漏极的耗尽层宽度有关，公式如下：

$$L_{\min}=A[r_j d(W_S W_D)^2]^{\frac{1}{3}} \tag{4-22}$$

式中，A 为经验常数；如果 r_j，W_S 和 W_D 为微米级而 d 为纳米级，那么 $A=0.88\text{nm}^{-1/3}$。

式（4-22）说明了小型元器件需要更浅的突变结（更小的 r_j）或者更薄的氧化层厚度（更小的 d）或者更小的耗尽层宽度（更低的电压或更重的掺杂度）。以上这些要求导致了很多难以控制的边界效应的产生。例如，如果氧化层在制造时使其厚度较薄一些，但同时偏置电压却没有相应地减小，那么随着氧化层场强的增加，就需要更好的氧化层来阻止边界效应产生。如果突变结深度减小，那么在元器件运行时就需要更好的控制，而且结电阻也会因为重叠部分变小而增加。为了控制结电阻，设计师们开发出了很多“自动校准接触”方案（Brews，1990；Einspruch and Gildenblat，1989），以便使源极和漏极电极更接近于栅极，以减小这些互连电阻。如果通过提高掺杂离子密度来减小耗尽层宽度，那么 MOSFET 的电荷驱动能力就会下降，因为开启电压提高了；也就是说，式

(4-9) 中的 Q_D 增加了，并导致 V_G-V_{TH} 减小了。因此，增加 V_{TH} 会使得电路的运行速度变慢。

提高掺杂离子密度导致的另一个结果是，沟道电导率会进一步下降，这主要是增加的掺杂原子的电子散射和增强的氧化层场强联合效应共同作用的结果，而反型层中的载流子接近于绝缘体-半导体界面，导致了界面散射的增加。这些效应同时还降低了元器件的电荷驱动能力，尽管在短沟道元器件中，电荷驱动能力只对于线性区域（即非饱和状态）比较重要，此时迁移率 μ 比饱和速度 v_{sat} 更有影响。

2. 热电子效应

另一个对 MOSFET 小型化产生限制的是指在小型元器件中大场强产生的直接影响。让我们从另一个角度来考虑一下，为什么更大的电压和场强会应用在小型元器件上。首先，根据式（4-14），如果电压增加会导致 τ_G 变小，至少变成 $V_G/L\leqslant F_{sat}\approx 5\times 10^4 V/cm$。如果 τ_G 按照上述方式减小，那么式（4-15）和式（4-16）中的 τ_D 和 τ_{INT} 也会同样减小。因此，通过提高速度饱和区域的电压可以增加元器件的响应速度。其次，对小型元器件的控制工艺还没有发展到随 L 同步对应比例减小的水平。因此，小型元器件参数还有很大变化空间。这样，就需要非对应比例的电压来确保电路中所有元器件正常运行，并克服增加的人为“干扰”。因此，为了提高电路运行速度，并应对工艺的变化，越小的元器件就需要越大的场强。

沿沟道方向的大场强产生的影响是指一小部分的沟道载流子拥有了足够的能量进入漏极附近的绝缘层。在硅基 P 沟道 MOSFET 中，高能量空穴被限制在了氧化层中，并导致漏极附近产生了氧化层正电荷，该电荷会使沟道长度变小，从而降低元器件的性能。在 N 型沟道 MOSFET 中，高能量电子进入氧化层后，会产生界面陷阱和氧化层消耗，从而造成栅-漏极短路（Pimbley 等，1989）。

为了解决上面遇到的问题，“漏极工程”应运而生，最常见的解决方案就是“轻掺杂漏极”(Chen，1990；Einspruch and Gildenblat，1989；Pimbley 等，1989；Wolf，1995)。在该方案的电路设计中，轻掺杂的扩展漏极被插入到沟道和漏极之间合适的位置。为了保持场强适中并减小场强的峰值，轻掺杂的扩展漏极在设计时必须尽量使漏极-沟道电压均匀地延伸。该方案的目标是使场强分布尽量平滑并处于接近于 F_{sat} 值，以便使高能量载流子保持最少。该方案的代价是导致了漏极电阻增加和增益降低。为了提高封装密度，该轻掺杂的扩展漏极可以垂直放置在栅极旁边，而不是水平放置在栅极下面，以便有效控制整个元器件的面积。

3. 薄氧化层

根据式（4-22），我们知道氧化层越薄，元器件就越小，元器件封装密度就越高。另外，减小电容性负载的响应时间，可以提高电荷驱动能力并增加输出阻

抗和跨导。在如何使氧化层变得更薄上还存在很多基本的限制。例如，氧化层场强存在一个绝缘层所能承受的最大值。我们原先以为 SiO_2 的本征崩溃电压可以达到 10^7V/cm 数量级，从而所能承受的场强可以达到 $\approx 2\times10^{13}$ 电荷/cm^2，这样该值足以使得场强限制变得无关紧要；但是，在实际制造过程中，MOSFET 氧化层的本征崩溃对场强产生的限制远小于由缺陷引起的渗漏或崩溃对场强产生的限制，因此控制制造工艺中的缺陷导致氧化层厚度最多只能减小至 5nm。

4. 掺杂离子的控制

随着元器件越来越小，元器件中掺杂物的精确位置越来越重要。器件在高温条件下运行时，掺杂离子可以移动。例如，源极和漏极掺杂离子可以进入沟道区域，并导致开启电压与掺杂离子的位置密切相关。类似的问题在隔离结构中同样存在，在隔离结构中，一个器件和另一个器件相互隔离（Pimbley 等，1989；Einspruch and Gildenblat，1989；Wolf，1995）。

为了有效控制上面这些热效应问题，元器件加工步骤必须仔细设计，以避免高温步骤对其产生的影响。这些设计可以通过计算机处理模型来实现和改进。但是，掺杂离子的运动是综合性的，其理论随着"快速热处理"的应用变得更为复杂，快速热处理包括瞬间热处理。因此，掺杂物响应过程不是一直处于稳定状态，而是瞬间的。计算机的瞬间响应模型是最原始的，这使得元器件的优化设计必须更注重于实践经验。

5. 其他限制

除了那些和 MOSFET 直接相关的限制外，在使用集成了很多元器件的小型 MOSFET 芯片时还有一些更明显的困难。例如，在前面已经提到的增加的延迟，该延迟是由于互连部分导致的，互连部分的长度随着芯片面积的和连接复杂度的不断增加而不断变长。MOSFET 中的那些会导致信号驱动能力下降的电容性负载会减慢电路的响应速度，因此需要附加电路来进行补偿。

另外一个关键就是需要将各个元器件相互隔离开（Brews，1990；Chen，1990，Einspruch and Gildenblat，1989；Pimbley 等，1989；Wolf，1995），以便在运行过程中通过寄生参数保持非耦合状态。当隔离结构越来越小时，元器件封装密度就越大，同时还会产生新的寄生参数。解决这个问题的方法就是在制造工艺中将电路印制在绝缘衬底上，即采用"绝缘硅"技术（Colinge，1991）；绝缘硅技术要取得成功还必须解决很多问题，例如衬底硅-绝缘层界面的电气特性以及绝缘层上面硅晶体的缺陷密度。

感谢

在此要感谢 R. D. Schrimf，并特别感谢为本书书稿进行审定的 S. L. Gilbert。

名词解释

沟道：MOSFET 中处于源极和漏极之间的导电区域。在增强型或者常闭型 MOSFET 中，沟道是指由栅极对少数载流子的吸引而形成的反型层。当栅极偏置超过开启电压值时，这些载流子形成了一个很薄的导电层，该导电层就被一层很薄的栅极氧化层与栅极隔离。在掩埋沟道型或者耗尽型或者常开型 MOSFET 中，即使栅极偏置为零，沟道也存在。因此，当栅极偏置电压非零时，栅极就可以用来提高沟道的电阻率。这样，这一类元器件就类似于 MESFET，都是基于多数载流子调节的。

栅极：MOSFET 的控制电极。栅极上的电压用来调节源极和漏极之间导电沟道的电阻率。

源极、漏极：MOSFET 的两个输出触点，通常与器件的衬底或主体形成 PN 结。

强反型：MOSFET 导通时的栅极偏置电压范围。当栅极偏置电压固定时，在低漏源偏置下，MOSFET 就等同于一个简单的栅极控制晶体管。当漏极偏置较大时，沟道电阻会随着漏极偏置的增加而增加，甚至达到电流饱和的状态，或者变得与漏极偏置无关。

衬底或主体：MOSFET 中处于源极和漏极之间、栅极之下的部分。栅极通过一层很薄的栅极绝缘层（通常为二氧化硅）和主体相互隔离。栅极调节主体的电导率，并在源极和漏极之间产生一个可由栅极控制的电阻。主体有时会添加直流电压来调节整体电路的运行状态。在有些电路中，主体电压会随着输入信号的变化而上下摆动，这一点会导致主体效应或反栅极偏置效应的产生；因此，要想得到可靠的电路响应，必须对这些效应进行有效控制。

亚阈值：MOSFET 处于截断状态时对应的栅极偏置电压。在该条件下，MOSFET 不是完全处于截断状态，而会产生一个渗漏电流，该电流必须进行有效控制，以避免产生电流误差和功率消耗。

开启电压：MOSFET 在截断和导通状态之间过渡的临界栅极电压值。

参考文献

下面所列的参考文献虽然不是本书内容的资料来源，但对读者来说还是比较有用的。

[1] Annaratone，M. 1986. *Digital CMOS Circuit Design*. Kluwer Academic，Boston，MA.

[2] Brews，J. R. 1981. *Physics of the MOS Transistor*. In *Applied Solid State Science*，*Supplement 2A*，ed. D. Kahng，pp. 1-20. Academic Press，New York.

[3] Brews，J. R. 1990. The Submicron MOSFET. In *High-Speed Semiconductor Devices*，ed. S. M. Sze，pp. 139-210. Wiley，New York.

[4] Chen, J. Y. 1990. *CMOS Devices and Technology for VLSI.* Prentice-Hall, Englewood Cliffs, NJ.

[5] Colinge, J. -P. 1991. *Silicon-on-Insulator Technology: Materials to VLSI.* Kluwer Academic, Boston, MA.

[6] Haznedar, H. 1991. *Digital Microelectronics.* Benjamin/Cummings, Redwood City, CA.

[7] Hollis, M. A. and Murphy, R. A. 1990. Homogeneous Field-Effect Transistor. In *Hign-Speed Devices*, ed. S. M. Sze, pp. 211-282. Wiley, New York.

[8] Einspruch, N. G. and Gildenblat, G. S., eds. 1989. *Advanced MOS Device Physics*, Vol. 18, VLSI Microstructure Science. Academic, New York.

[9] Malik, N. R. 1995. *Electronic Circuits: Analysis, Simulation, and Design.* Prentice-Hall, Englewood Cliffs, NJ.

[10] Nicollian, E. H. and Brews, J. R. 1982. *MOS Physics and Technology*, Chap. 1. Wiley, New York.

[11] Pearton, S. J. and Shah, N. J. 1990. Heterostructure Field-Effect Transistor. In *High-Speed Semiconductor Devices*, ed. S. M. Sze, pp. 283-334. Wiley, New York.

[12] Pierret, R. F. 1990. *Field Effect Devices*, 2nd ed., Vol. 4, *Modular Series on Solid State Devices.* Addison-Wesley, Reading, MA.

[13] Pimbley, J. M., Ghezzo, M., Parks, H. G. and Brown, D. M. 1989. *Advanced CMOS Process Technology*, ed. N. G. Einspruch, Vol. 19, *VLSI Electronics Microstructure Science.* Academic Press, New York.

[14] Sedra, S. S., K. C. 1991. *Microelectronic Circuits*, 3rd ed. Saunders College Publishing, Philadepia, PA.

[15] Shoji, M. 1988. *CMOS Digital Circuit Technology.* Prentice-Hall, Englewood Cliffs, NJ.

[16] Wolf, S. 1995. *Silicon Processing for the VLSI Era: Volume3——The Submicon MOSFET.* Lattice Press, Sunset Beach, CA.

备注

上面列出的参考文献为本书中的内容给出了更为详细的参考。尤其是，Annaratone (1986) 和 Shoji (1988) 的著作中提供了很多关于元器件和电路性能的详细描述；Chen (1990)、Pimbley (1989) 和 Wolf (1995) 的著作中主要介绍了许多关于元器件运行的技术细节及其对元器件的影响；Haznedar (1991)、Sedra &Smith (1991) 和 Malik (1995) 的著作中提供了许多关于电路的信息；Brews (1981) 和 Pierret (1990) 的著作中主要对元器件伏安特性曲线的由来以及在所有偏置电压范围内元器件性能的进行了讨论。

第5章 集成电路

Tom Chen

5.1 引言

晶体管及其在超大规模集成（Very Large Scale Integrated，VLSI）电路中的制造工艺是一项伟大的发明，这项发明使得现代计算技术成为现实。集成电路（Integrated Circuit，IC）的规模从20世纪60年代初的在一个很小的硅片上集成少量晶体管已经发展到了在一个较大的单硅衬底上可以集成400万个晶体管的水平。目前，应用在集成电路中的晶体管类型主要是金属-氧化物-半导体型（Metal Oxide Semiconductor，MOS）晶体管。集成电路技术在20世纪80年代及80年代后期得到了快速发展，其主要推动力是技术的集成，即MOS晶体管外形尺寸的小型化。MOS晶体管外形尺寸的主要衡量标准是MOS晶体管中导电沟道的长度。晶体管越小，硅衬底单位面积上集成的晶体管数量就越多，因此集成电路的集成度就越高，从而晶体管的转换速度就越快。目前，我们不仅可以提高单位面积上晶体管的数量，还可以扩大衬底芯片的面积。随着晶体管尺寸越来越小以及芯片规模越来越大，晶体管的电荷驱动能力会下降，而且互连寄生参数（互连电容和电阻）也会增加。因此，整个VLSI系统的设计必须非常细致，以满足未来速度上的要求。常见的设计要素包括最佳门电路设计、晶体管尺寸、时钟脉冲相位差、合适的定时预算以及互连寄生参数的实际模型。

5.2 高速电路设计技巧

现代VLSI元器件中都包含许多大规模的单元电路如存储块、数据分支算法块，以及许多基本MOS逻辑门电路（如反相器和NAND/NOR门电路）。互补金属氧化物半导体（CMOS）电路就是应用最广泛的逻辑系列之一，这主要是因为CMOS电路具有低功耗和高噪声容限的特点；其他逻辑系列还包括NMOS和PMOS逻辑电路。由于CMOS电路使用最广泛，因此我们接下来只讨论CMOS逻辑电路。本书讨论的高速电路设计技巧对其他逻辑系列同样适用。

高速电路中VLSI元器件的优化可以在系统级的水平进行，如同在电路和逻辑水平一样。为了在给定技术条件下的电路和逻辑水平上实现最大运行速度，必

须合理设置逻辑门电路中每个晶体管的尺寸，以便得到最佳输出负载。如果输出负载非常大，就需要一连串尺寸呈几何增加的驱动器。逻辑门电路中晶体管的尺寸还与晶体管作为负载有关。作为负载时，晶体管由其前置门电路驱动。

1. 门限设计的优化

为了优化门限设计，让我们先了解一下简单 CMOS 反相器的性能，如图 5-1 所示。

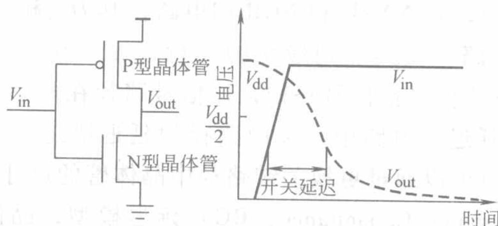


图 5-1 简单反相器中的门电路延迟示例

门电路延迟是指输入过渡电压和输出过渡电压在 50% 处的时间差。因此，反相器门电路延迟可以描述如下：

$$T_d = C_l(A_n/\beta_n + A_p/\beta_p)/2$$

式中，C_l是指反相器的负载电容；β_n和β_p分别为 N 型晶体管和 P 型晶体管的前向电流增益，它们与晶体管的沟道宽度成正比，而与晶体管的沟道长度成反比；A_n和A_p为给定电压下的过程相关参数，由下式定义：

$$A_n = [2n/(1-n) + \ln((2(1-n) - V_0)/V_0)][V_{dd}(1-n)]$$

$$A_p = [-2p/(1+p) + \ln((2(1+p) - V_0)/V_0)][V_{dd}(1+p)]$$

式中，$n = V_{thn}/V_{dd}$；$p = V_{thp}/V_{dd}$。其中，V_{thn}和V_{thp}分别为 N 型晶体管和 P 型晶体管的栅极开启电压。

上面的公式没有将输入信号斜率考虑在内；否则，公式将会变得非常复杂。对于很多复杂的 CMOS 门电路来说，为了使用反相器延迟模型，必须构建一个等效的反相器结构，以反映 CMOS 门电路中 P 型树状结构和 N 型树状结构的有效强度。

在实际应用中，对 CMOS 门电路延迟的处理很简单。逻辑门电路的延迟可以分为两部分：本征延迟 D_{ins}和负载相关延迟 D_{load}。门电路本征延迟由门电路内部特征决定，这些特征包括制造技术、门电路结构和晶体管尺寸；负载相关延迟是门电路输出时总体负载电容的函数。门电路总体延迟可以表述如下：

$$T_d = D_{ins} + C_l^* S$$

式中，C_l是指总体负载电容；S是指门电路驱动能力因子；$C_l^* S$代表了门电路的负载相关延迟。

在大多数采用前沿亚微粒技术的CMOS电路中，门电路的总体延迟由负载相关延迟决定。对于采用现代亚微型CMOS技术（0.5μm级）的反相器来说，其本征延迟D_{ins}的变化范围为0.08～0.12ns，S的变化范围为0.00065～0.00085ns/fF，具体数值取决于制造技术的细节和晶体管的最小尺寸。对于其他更加复杂的门电路来说，如NAND和NOR门电路，其D_{ins}和S的值通常会更大。

为了优化VLSI电路，使其达到最快的运行速度，我们必须确定“关键路径”。在电路中，关键路径是指信号路径，该信号路径在从主要输入端至主要输出端之间具有最长的延迟。电路中，关键路径的延迟决定了电路的最高运行速度。关键路径的延迟也可以通过更换关键路径中晶体管的尺寸来减小。通过使用集总电阻-电容（Resistance-Capacitance，RC）延迟模型，晶体管的尺寸问题可以变成一个最优化问题，并使路径延迟和晶体管尺寸之间产生紧密的联系。这个最优化问题很容易得到解决，但是解决方法相比SPICE仿真的结果经常存在20%～30%的偏差。在实际门电路延迟模型中，我们必须考虑次要变量如输入信号斜率，这表明路径延迟和晶体管尺寸之间的关系还不是那么紧密。上述详细的分析涉及到了很多精细的晶体管尺寸计算方法，其中之一就是采用起源法来寻找一个最优化的解决方案，该方法已经取得了一定的成效。

2. 高速电路设计中的时钟及时钟设计

大多数现代电子系统都是同步系统，而时钟就是同步系统中的中央同步设置单元，该单元通过各种计算步骤来调整理想系统的运行。锁存器经常用来在每个时钟周期的末端获取输出数据。图5-2给出了典型同步电路的例子。在该同步电路中，随机逻辑串作为计算模块，锁存器作为同步设置器件。当存在反馈时（见图5-2），该电路就变成了时序电路。

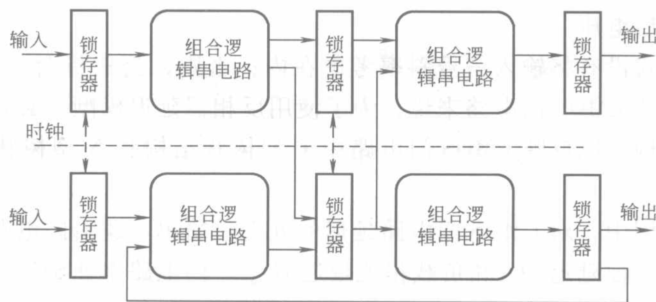


图5-2 带有锁存器和组合逻辑串的典型同步电路示例

锁存器也称为“寄存器”或“触发器”。锁存器获取数据的方法与时钟信号触发数据的方式有关。通常，触发器分为电平触发和边沿触发，其中，电平触发

还可以根据触发极性为正或负来进一步细分。数字电路的性能通常由电路可以运行的最高时钟频率决定。对于同步数字电路来说，正常工作时，通过任何一个组合逻辑串的最大延迟必须小于时钟周期。因此，在高速电路的设计中必须做到以下几点：

1）划分整个系统，以便使所有组合逻辑串的延迟尽可能均匀。

2）规划组合逻辑串电路，以便使电路中关键路径的延迟最小，且低于理想的时钟周期。

3）采用强壮的时钟方案，以确保整个系统与竞争条件无关，并具有最小的、可容忍的时钟脉冲相位差。

上面几点中的第一点不属于本节讨论的范围，而第二点已经在前面的内容中讨论过了。

第三点中的竞争条件通常会随着电平触发锁存器的使用而出现。图5-3给出了一个典型的基于电平触发锁存器的同步系统例子。由于时钟分布网络的延迟（如缓冲器和互连部分的电容性寄生参数），由该分布延迟导致的时间差通常称为“时钟脉冲相位差”；在图5-3中，时钟脉冲相位差由延迟元器件进行模拟。为了使系统正常运行，在正极边沿处，每个锁存器应该从前一个时钟周期获取输入数据。但是，如果时钟脉冲相位差标识为倾斜时钟clk″，这说明了$Q1$至$D2$的延迟变得足够小，以至于$D1$不仅仅可以在clk上获取，还可以在clk′上获取。为了防止由严重的时钟脉冲相位差导致竞争条件的产生，可采用如下所述的方法：

1）将锁存器的构成元器件由电平触发改为边沿触发或者伪边沿触发，如二相、互不重叠时钟锁存器。

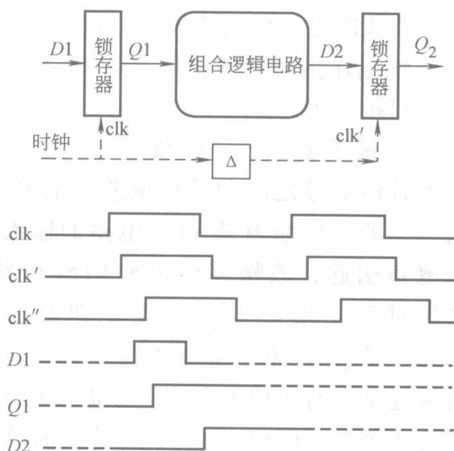


图5-3 由一个严重的系统时钟脉冲相位差导致的竞争条件示例

2）重新合成系统，以平衡不同组合逻辑串电路的关键路径延迟。

3）减小时钟脉冲相位差。

如果不给予足够的重视，时钟脉冲相位差还会导致其他类型电路故障的产生。在动态逻辑电路中，严重的时钟脉冲相位差会导致不同的功能模块处于不同的预先充电或评估阶段。在 VLSI 系统中，时钟脉冲相位差会吞噬掉很多宝贵的时钟周期时间。因此，在高速电路设计中减小时钟脉冲相位差是一件非常重要的事情。

如前所述，时钟脉冲相位差是由于时钟分布网络的不均衡导致的，这种不均衡产生的原因包括最近的时钟驱动器距离差异、不同的功能模块由不同的时钟驱动器驱动（该驱动器具有不同的驱动能力）、相同模块的温度差异、相同模块上由于工艺参数产生的元器件特征差异等等。通常，我们采用两种方法来使时钟脉冲相位差最小化。第一种方法就是调整时钟信号的分布方式。时钟分布网络的几何形式是一个很重要的特征属性。根据系统运行的类型，图 5-4 给出了许多很受欢迎的分布网络拓扑的结构。在这些拓扑结构中，H 树形结构代表了最小的时钟脉冲相位差，因此，H 树形结构广泛应用于高性能系统中。

第二种方法是通过添加片上电路来迫使两个不同功能模块的时钟信号进行校准，或者迫使片上时钟信号根据整体时钟信号在系统级水平上进行校准。基于这种目的而且应用广泛的电路包括：锁相环（Phase-Locked Loop，PLL）电路和延迟锁环（Delay-Locked Loop，DLL）电路。图 5-5 给出了一个简单的锁相环电路例子。一个简单的锁相环（PLL）电路由 4 部分组成：相位检波器、电荷泵、低通滤波器和电压控制振荡器（Voltage Control Oscillator，VCO）。其中，相位检波器接受参考时钟 CLK_ref 和错位时钟 CLK_out，并通过比较两者之间的相位差来判断是否对电荷泵进行充电或放电；低通滤波器负责将参考频率和错位频率之间的相位差转换成电压值。该电压值随后被输入 VCO 来减小参考频率和错位频率之间的差异，直到它们之间相互锁定。

输出抖动是 PLL 电路中最重要的设计参数之一。在电路中，输出抖动的标志是输出时钟相位与参考时钟信号之间的随机偏差。重要的峰-峰抖动会有效减小时钟周期。产生输出抖动的主要原因是 VCO 电路的输入端噪声。其余的抖动可能由电源保护装置的噪声引起，该噪声在高速 VLSI 电路中普遍存在。另外，PLL 电路的输入信号采集时间（几 μs 内）通常比理想的要长，这主要是由于 VCO 电路响应时间的缘故。在由时钟分布网络的不均匀导致的时钟脉冲相位差典型情况中，错位的时钟通常具有正确的频率，而需要校正的只是时钟信号的相关相位，因此，就没必要再添加 VCO 电路了。另外，简单的延迟逻辑电路就可以用来校准时钟信号相位。

这种简单的相位校准电路被称为“延迟锁环（DLL）电路”。通过将 VCO 替

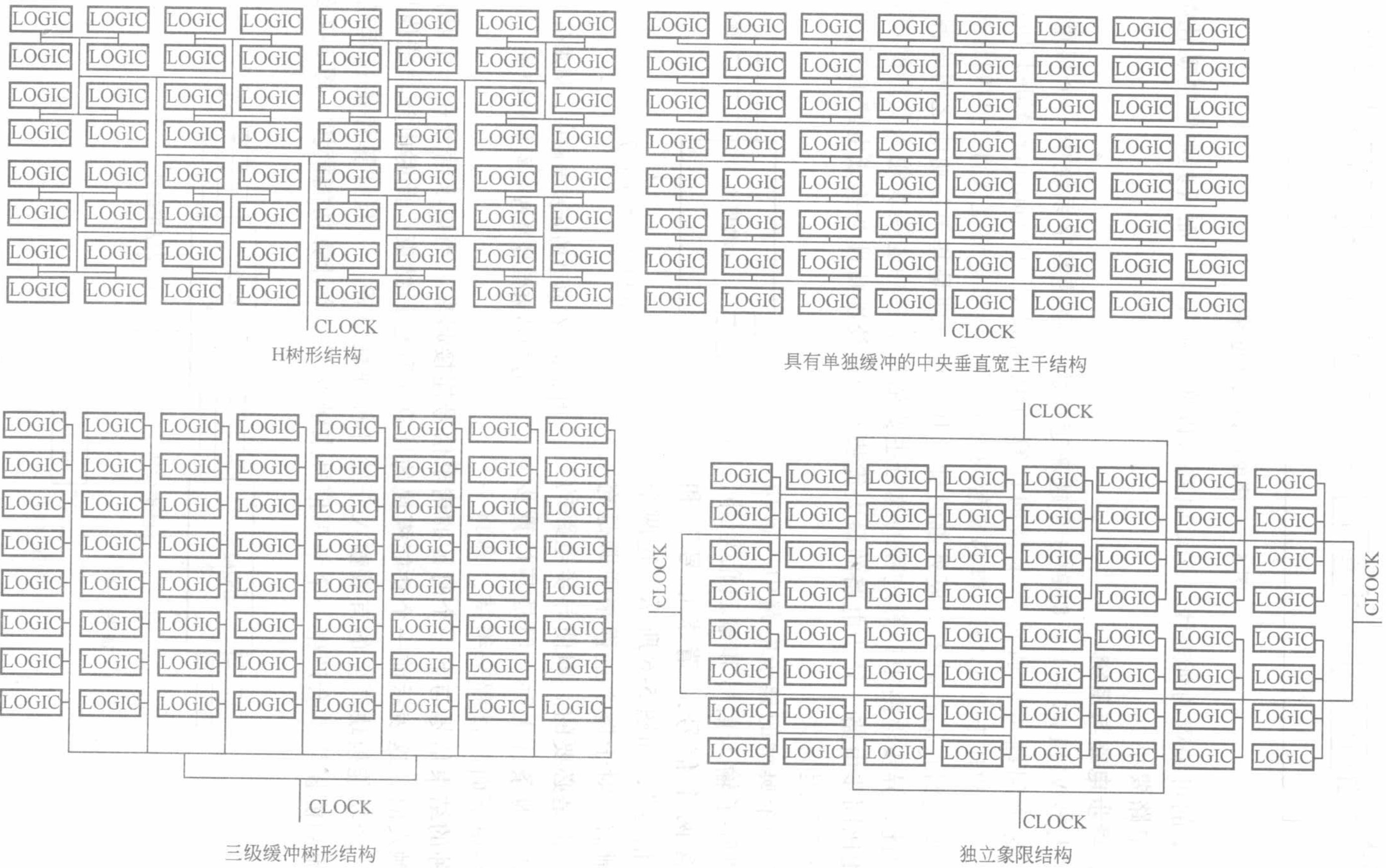


图 5-4　各种时钟分布结构

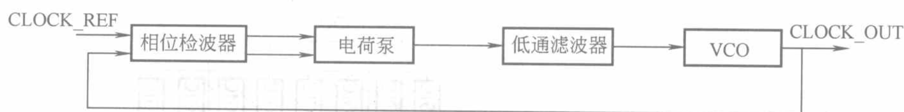


图 5-5 简单的锁相环电路

换成简单的可编程的延迟线，DLL 电路可以变得更简单，而且其电路抖动比类似的 PLL 电路要小。

3. 异步电路与系统

在大型 VLSI 芯片上，时钟分布越来越成为了高速数字系统中一个很重要的问题。这个问题可以通过艺术级的计算机辅助设计（Computer Aided Design，CAD）工具和片上 PLL/DLL 电路来解决。不过，异步电路受关注的时间比较晚。一个异步电路在运行时不需要外部时钟，其工作原理是基于各功能之间的握手原则。因此，异步电路中计算过程的执行主要依赖于各功能模块输入变量的预先状态。相比同步电路，异步电路的最大优势在于其校准特性与构成电路的元器件和信号互连延迟无关。

在一个典型的异步电路中，各功能模块除了输入信号和输出信号外，还包含两个信号:“请求信号”和“结束信号”，如图 5-6 所示。这两个二进制信号对于握手过程来说既是充分的，也是必要的。即使异步电路的速度和其他条件无关，但是其计算步骤仍然由下面的过程来维持，即将一个时钟的结束信号与下一个时钟的请求信号连接起来。当请求信号在一个功能模块中触发时，这表示前一个功能模块的计算过程已经结束，此时电流功能模块利用前一个功能模块产生的有效输入开始计算评估过程。一旦计算评估过程结束，该电流功能模块就会产生一个结束信号来触发下一个计算功能模块。图 5-7 给出



图 5-6 在典型的异步电路中，请求信号和结束信号是两个附加的信号

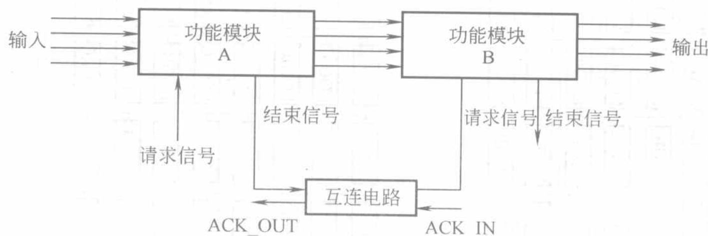


图 5-7 异步电路系统中的通信协议示意图

了一个通信协议的示意图，其中模块 A 和模块 B 通过管线相互连接。

为确保异步电路的功能与各单独模块的速度无关，如果各功能模块已经完成了电流计算过程，那么该功能模块的请求信号只能由某个模块单独触发。否则，该电流计算过程就可能会在输入的计算请求下重新进行。为了防止这种情况的发生，就需要一个互连模块，该互连模块在电流功能模块和前一个功能模块之间产生一个“确认信号”。对于前一个功能模块来说，一个动态的确认信号说明电流功能模块已经准备好从前一个功能模块接受新的数据了。这种具有请求信号和确认信号的双向通信协议如图 5-7 所示。其中，互连电路是异步电路中特有的部分，通常称为“*C*-单元”。图 5-8 给出了 *C*-单元的设计示意图。

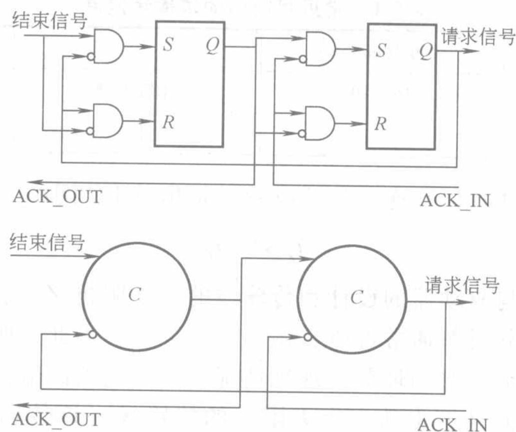


图 5-8 *C*-单元设计示意图

近几年，人们为了将异步电路运用到实际应用中付出了很多努力，许多微处理器的完全异步设计就充分证明了其商业上的可行性。但是在异步电路中还需要注意很多要素，例如可接受的硅片数量、电源效率以及相比同步电路的性能。

4. 互连寄生参数及其在高速电路设计中的影响

在有源电路中，片上互连部分会产生寄生电容和电阻，并成为电路的负载。这些寄生负载在早期的 IC 中几乎没有影响，因为本征门电路延迟决定了整体的门电路延迟。随着 VLSI 的大规模发展，门电路本征延迟大幅度下降了，但是，互连寄生参数并没有相应地改善，而且导线电阻越来越大，导致由互连负载寄生参数产生的延迟逐渐变成了整体门电路延迟中的主要部分。随着运行速度达到数百万赫兹，这个问题就更加恶化了，传统的集总 RC 模型就不再准确了。人们建议在这种传统的 RC 模型中添加一个接地电阻和一个电感来对电路进行校准。RLC 互连模型包含了非均匀的初始条件，其响应波形可能是可变的；这种模型可能更加准确，因为电感的存在会减缓电流增加的速度，从而增加了信号跃迁的

时间。当运行速度进一步提高时，信号的上升时间就远比信号从 A 点传输至 B 点所需的时间少，此时就需要使用传输线模型。因此，片上互连部分经常被模拟成微带模型。

传输线的特性由其介电常数和导磁率决定。表 5-1 给出了应用在 VLSI 中的一些常见材料的信号传输速度。根据经验，传输线现象在满足如下条件时变得更加明显。

$$t_r < 2.5 * t_f$$

式中，t_r是指信号的上升时间；t_f是指信号的传输时间，该传输时间等于给定材料中的互连长度除以信号传输速度。

表 5-1 常见材料中的传输线速度

	速度/(cm/ns)		速度/(cm/ns)
聚酰胺	16 ~ 19	环氧玻璃	13
SiO_2(二氧化硅)	15	铝	10

当满足如下条件时，互连部分可以看做是集总 RC 网络。

$$t_r > 5 * t_f$$

信号上升时间与驱动器的设计和传输线的特征阻抗 Z_0 有关。在 MOS IC 中，传输线接收端的负载设备通常可以看做是一个开路。因此，驱动器的设计是高速电路设计中一个非常重要的环节。理想情况下，驱动器的输出阻抗与传输线的输出阻抗相互匹配。但是，驱动一个无限长的传输线（该传输线的输出阻抗低于特征阻抗）会增加驱动器的建立时间（由于过剩的激振），因此必须尽量避免发生。接收端过剩的激振会导致负载产生不良的转换。假设 MOS 晶体管的开启电压为 0.6 ~ 0.8V，为了确保避免不良转换的发生，驱动器的输出阻抗必须至少是传输线特征阻抗的 1/3。当输出阻抗高于传输线特征阻抗时，就需要使用信号的多个波形跳跃来变换负载。为了确保在变换负载时只需要一个波形跳跃，驱动器的输出阻抗必须在传输线特征阻抗的 60% 以内。

对于有损耗的传输线来说（该损耗由片上互连部分的寄生电阻产生），在信号传输过程中就需要在传输线上的任意一点上使用指数衰减传输函数，衰减速度和互连部分的单位阻抗成正比。当工作频率超过某个值时，片上传输介质就会产生“表面效应”。在表面效应中，时变电流就会集中到导体的表面附近，因此，传输介质的单位阻抗就会大幅度地增加。

名词解释

专用集成电路（**ASIC**）：应用于特定用途的器件。

专用标准产品（ASSP）：应用于某一领域的器件，如图像和视频处理电路元器件。

异步系统：运行过程由所有必要输入变量的预先状态通过握手协议来驱动的系统。因此，不需要外部时钟。

C-单元：应用在异步系统中作为互连单元的电路。这种电路的作用是为了使两个功能模块之间握手通信协议变得更容易。

时钟脉冲相位差：一个芯片或系统中不同部分的两个信号之间的相位差，主要是由于介质和网络的分布不均衡引起的。

互补金属氧化物硅半导体（CMOS）：目前非常受欢迎的一种集成电路。

关键路径：逻辑模块中从主要输入端至主要输出端的信号路径，在该路径上延迟最长。

延迟锁环（DLL）：类似于PLL，其抖动抑制能力更强。

数字信号处理器（DSP）：根据信号处理规则，应用于数学程序的处理器。

现场可编程门阵列（FPGA）：一种可以根据电路需求，并通过定制相应的程序来进行现场设计的元器件。

H树形结构：树形时钟分布拓扑结构，其形状类似于字母“H”。和其他分布拓扑结构相比，该结构具有最小的时钟脉冲相位差。

锁相环（PLL）：可以检测两个信号之间相位差并降低当前相位差的电路。

可编程逻辑器件（PLD）：可以很容易地为特定用途进行量身定制的一系列IC产品。

SPICE：一个很受欢迎的电路级仿真程序，用来对电路性能进行详细分析。

同步系统：运行过程被划分为单位时段的系统，该单位时段由中心时钟信号定义；系统信号发送由时钟边沿触发。

参考文献

[1] Bakoglu, H. B. 1991. *Circuits, Interconnections, and Packaging for VLSI.* Addison-Wesley, Reading, MA.

[2] Dill, D. L. 1989. *Trace Theory for Automatic Hierarchical Verification of Speed-Independent Circuits.* MIT Press, Cambridge, MA.

[3] Gardner, F. M. 1979. *Phaselock Techniques*, 2nd ed. Wiley, New York.

[4] Jeong, D. et al. 1987. Design of PLL-based clock generation circuits. *IEEE J. Solid-State Circuits* SC-22 (2): 255-261.

[5] Johnson, M. and Hudson, E. 1988. A variable delay line PLL for CPU-coprocessor synchronization. *IEEE J. Solid-State Circuits* (Oct.): 1218-1223.

[6] Meng, T. H. 1991. *Synchronization Design for Digital System.* Kluwer Academic, Norwell, MA.

[7] Rosenstark, S. 1994. *Transmission Lines in Computer Engineering.* McGraw-Hill, New York.

[8] Sapatnekar, S., Rao, V., and Vaidya, P. 1992. A convex optimization approach to transistor sizing for CMOS circuits. *Proc. ICCAD*, pp. 482-485.

[9] Wang, X. and Chen, T. 1995. Performance and area optimization of VLSI systems using genetic algorithms. *Int. J. of VLSI Design* 3 (1): 43-51.

[10] Weste, N. and Eshraghian, K. 1993. *Principle of CMOS VLSI Design: A Systems Perspective*, 2nd ed. Addison-Wesley, Reading, MA.

备注

关于 VLSI 的设计和各种设计要素，读者可以参考各种相关的优秀书籍，本部分列出了其中的三本：Mean 和 Conway 的《*Introduction to VLSI Systems*》、Glasser 和 Dobberpuhl's 的《*The Design and Analysis of VLSI Circuits*》以及 Geiger 的《*VLSI Design Techniques for Analog and Digital Circuits*》。*IEEE Journal of Solid-State Circuits* 中提供了关于新奇的和高性能 VLSI 器件的最新发展信息。

关于 PLL 和 DLL 电路的最新应用信息，读者可以参考以下会议资料："Proceedings of International Solid-State Circuits Conference"、"*Symposium on VLSI Circuits*" 以及 "*Custom Integrated Circuit Conference*"。

关于 VLSI 互连部分及其传输线特性的仿真，读者可以参考以下会议资料："*Proceedings of Automation Conference*" 和 "*International Conference on Computer-Aided Design*"。另外，"*IEEE Transactions on CAD*" 也是关于该主题的优秀资料。

第6章 集成电路设计

SamuelO. Agbo EugeneD. Fabricius

6.1 引言

集成电路（IC）根据其集成度可以分为：小规模集成（Small Scale Integration，SSI）、中规模集成（Medium-Scale Integration，MSI）、大规模集成（Large Scale Integration，LSI）和超大规模集成（Very Large Scale Integration，VLSI）电路；也可以根据制造工艺进行分类（如双极性N型金属氧化物半导体（NMOS）、互补金属氧化物半导体（CMOS）等等）。因此，集成电路的设计必须基于SSI、MSI、LSI或VLSI规模。其中，数字SSI和MSI基本上由门电路和组合门电路构成。关于数字SSI和MSI的设计在6.3节中将有详细介绍，其大部分内容是关于标准门电路的。这些标准门电路通常都被设计成具有很大的噪声容限、输出端数以及负载电流能力，以便拓展各种功能。

原则上，无论数字集成电路如何复杂，基本门电路在任何数字电路的设计中都是最基本的。而实际上，当基本门电路和MSI电路（如触发器、寄存器、加法器等等）应用到LSI或VLSI设计中时，就需要对这些电路进行改进。例如，在同一块芯片上互连的电路可以被设计成具有低噪声容限、低负载驱动能力以及更小的逻辑电压摆动等特点的电路；这样设计的好处是使电路具有更低的功耗、更高的电路集成度和更高的可靠性。从另一种角度来说，LSI和VLSI的设计过程采用了很多方法，LSI和VLSI的设计不再仅仅是基于互连或SSI和MSI电路的改进了。LSI和VLSI的具体设计方法将在接下来的内容中介绍。

6.2 IC设计过程概述

设计一个集成电路所要求的工作量与集成电路的复杂程度有关。设计需要的工作量可以是单个设计人员几天时间的工作量到一个设计组人员几个月工作量，用户定制的复杂集成电路设计需要最多的工作量。相比之下，半定制设计的LSI和VLSI利用已有的设计如标准单元和门阵列则需要较少的设计工作量。

IC 设计过程可以分为许多步骤，图 6-1 给出了这些步骤的一个示例。其中，第 1 步是各子系统及其互连部分的设计（如标准单元、门阵列和用户定制的子电路）。系统规划的设计从第 3 步的平面方案设计开始，不包括各独立晶体管和元器件的规划，但是涉及到几何阵列和子系统的互连部分。第 4 步是子系统的电路设计。第 2 步和第 5 步分别包含了系统仿真和电路仿真，这两步会对第 1 步或第 4 步进行改进。

在此，我们将主要讨论第 1 步的系统设计和第 4 步的子系统电路设计。第 7 步的制造工艺之后还包括很多任务，如掩模制造、过程仿真、晶片制作以及测试等等。总体来说，平面方案设计是总体规划的一部分。对于大规模 IC 来说，电路布置图设计常常与系统及电路设计密切相关。

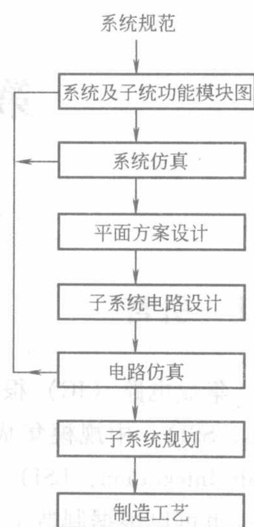


图 6-1 IC 设计的各个步骤

6.3 IC 设计总则

IC 芯片上各元器件的布置及实际排列是一个包含了很多设计评比方案的设计过程。电路设计常常会影响到整体规划，反过来，整体规划也会影响电路的设计，尤其是在 LSI 和 VLSI 中。因此，非常有必要给出一些 IC 设计的设计总则，如下所述：

1）电路元件：芯片面积对于 IC 设计至关重要。常见的 IC 元器件包括晶体管、电阻和电容。电感对于面积要求来说是非常不经济的，因此在 IC 中几乎不使用，除了某些微波电路。晶体管所占的面积很小，因此在 IC 中的使用率很高。电阻所占的面积随电阻值的增加而增加。IC 中使用的电阻阻值范围为 50Ω ~ 100kΩ。电容对面积是非常敏感的，因此，其值通常被限制在 100pF 以内。

2）隔离区域：通常不同的元器件会被放置在不同的隔离区域中。但是，基于芯片面积的经济考虑，隔离区域的数量必须尽量少，因此可以在每个隔离区域尽可能地放置多个元器件。例如，许多电阻可以共享一个隔离区域。

3）设计规则：指定了最小特征尺寸的几何设计规则，各特征尺寸之间必须根据给定 IC 设计流程保留空隙。

4）功耗：芯片规划必须考虑到足够的功耗，并避免芯片过热或者在芯片上出现热点。在低功耗电路（如 CMOSIC）中，元器件的尺寸由平版印制的约束条件决定。在具有可评估功耗的电路中，元器件的尺寸由热效应的约束条件决定；

该条件下的元器件尺寸可能比平版印制约束条件决定的元器件尺寸大很多。

图 6-2 阐述了根据功率密度条件来决定元器件尺寸的原理。元器件处于衬底表面附近，因此从元器件至衬底的热流尽管事实上是三维的，现在就变成一维的了。假设无限散热片的温度为环境温度 T_A，衬底厚度为 ΔX，热导率为 α，元器件表面面积为 A，温度为 $T_A+\Delta T$，那么热流散热的速率为 dQ/dt，如式（6-1）所示：

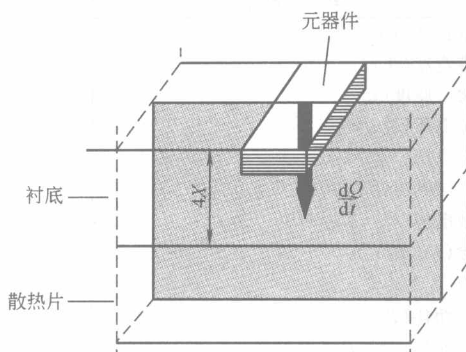


图 6-2　IC 元器件中的热流

$$\frac{dQ}{dt}=\alpha A\left(\frac{\Delta T}{\Delta X}\right) \tag{6-1}$$

元器件的功率密度或单位面积上功率的分布为

$$\frac{P}{A}=\alpha\left(\frac{\Delta T}{\Delta X}\right) \tag{6-2}$$

1. 元器件缩小

IC 的设计正日益朝着更小尺寸的方向发展（尤其是 VLSI）。缩小技术使得电路的设计可以随着技术的革新而不断向更小尺寸的方向发展。缩小分为两种基本的方式：全面缩小和常压缩小，如下所述。

1）全面缩小：所有元器件的尺寸（包括表面方向和垂直方向的）以及所有电压都随相同的缩放比例 S 减小。

2）常压（ConstantVoltage，CV）缩小：所有元器件的尺寸（包括表面方向和垂直方向的）随相同的缩放比例 S 减小，但是所有的电压不进行缩小，其值保持在和晶体管-晶体管逻辑（Transistor-Transistor Logic，TTL）电源电压及逻辑电压相一致的水平。

元器件尺寸的缩放对其他元器件参数也会产生影响。全面缩小会保持电场强度不变，因此，当元器件尺寸变小时，其他参数不会出现恶化，但是不能确保 TTL 电压的兼容性。表 6-1 比较了两种缩小方式对元器件参数产生的影响。通常，我们采用折中的方法即对所有内部电路使用全面缩小，但同时在芯片的输入/输出（I/O）脚上也要保持 TTL 电压的兼容性。

尽管许多缩放关系对于 MOS 场效应晶体管（MOSFET）和双极性 IC（Keyes，1975）来说很普遍，但是表 6-1 中的缩放关系应用到 MOSFET 时要求更加严格。双极性掺杂工艺不像 MOSFET，它不受氧化层崩溃的限制。因此，原则

表 6-1　全面缩小和常压缩小对 IC 元器件参数产生的影响

参　数	完全缩放	CV 缩放
沟道长度(L)	$1/S$	$1/S$
沟道宽度(W)	$1/S$	$1/S$
氧化层厚度(t_{OX})	$1/S$	$1/S$
电源电压(V_{DD})	$1/S$	1
开启电压(V_{TD})	$1/S$	1
氧化层电容(C_{OX}, C_{SW}, C_{FOX})	S	S
栅极电容($C_g = C_{OX}WL$)	$1/S$	$1/S$
跨导(K_n, K_p)	S	S
电流(I_D)	$1/S$	$1/S$
密度功耗(P)	$1/S^2$	S
单位面积功耗(P/A)	1	S^3
封装密度	S^2	S^2
传播时延(t_p)	$1/S$	$1/S^2$
功率-时延乘积(Pt_p)	$1/S^3$	$1/S$

上，双极性处理工艺可以使元器件进一步小型化。但是，双极性缩放更加复杂，其中一个原因就是用来开启双极性结晶体管的结电压不会随着元器件尺寸的减小而下降。

2. 几何设计规则

几何设计规则中指定了 IC 上的最小元器件尺寸、特征尺寸间的最小间隔以及特征尺寸间的最大校准误差。这些规则通常与制造工艺和设备有关。例如，2μm 工艺的设计规则不一定适合 0.5μm 工艺。设计规则应尽量避免出现重大的错误，例如由于特征尺寸间过大的校准误差而导致的短路，或者由于金属导电路径或多晶硅导电路径过窄而导致的开路。

在 IC 设计中，适用于各种工艺以及可随工艺收缩最小几何尺寸的常见设计规则是必要的。常见设计规则还包括其他优势：由于步骤和规则更少从而提高了设计效率、自动转到最后的规划、规划原则和电气原则检查、仿真以及验证等等。常见设计规则中的 Mead-Conway 方法定义了一个可升级的和工艺相关的参数 λ，该参数是晶片上元器件指定位置处的最大特征尺寸校准误差，或者不同掩模上两个特征尺寸之间的最大校准误差的半值。表 6-2 给出了 Mead-Conway 可升级的 NMOS 设计规则（Fabricius，1990）。

CMOS IC 中既包含 NMOS 也包含 PMOS。如果 NMOS 是基于 P 型衬底工艺，那么 PMOS 就是基于 N 型电位阱工艺，反之亦然。如果添加了 N 型电位阱、P 型电位阱或者双槽工艺，那么 CMOS 工艺就类似于 NMOS 工艺了，甚至更加复杂。表 6-3 给出了 Mead-Conway 可升级的 CMOS 设计规则（Fabricius，1990）。在表 6-3 中，器件的尺寸由 λ 的倍数表示，设计规则由 Southern California 大学信息科学学院的 MOS 执行系统（MOS Implementation System，MOSIS）指定，该规则的细节可参考 MOSIS 手册（USCISI，1984，1988）。

表 6-2　MOS 执行系统（MOSIS）中的 NMOS 设计规则

掩模类型	特征尺寸	大小(λ的倍数)
N^+扩散	扩散宽度	2
	扩散间距	3
注入掩模	注入区域与栅极的重叠宽度	2
	注入区域与栅极的间距	1.5
隐埋接触掩模	与有源元器件的接触	2
	扩散重叠	1
	触点与聚合物的间距	2
	触点与扩散区域的间距	2
聚合物(Poly)掩模	聚合物宽度	2
	聚合物间距	2
	聚合物与扩散区域的间距	1
	扩散区域外的栅极扩展	2
	聚合物边缘的扩散	2
接触掩模	触点宽度	2
	触点与扩散区域的重叠	1
	触点与聚合物的重叠区域	1
	触点之间的间距	2
	沟道上的触点	2
	触点与金属的重叠区域	1
金属掩模	金属宽度	3
	金属间距	3

表 6-3　MOSIS 中的 CMOS 设计规则

掩模类型	特征尺寸	大小(λ的倍数)
N型电位阱和P型电位阱	阱宽度	6
	阱间间距	6
N^+,P^+型有效扩散或注入	有效扩散宽度	3
	有效扩散区域间的间距	3
	源极/漏极至阱边缘的间距	6
	衬底/阱触点	3
	有效扩散区域至阱边缘的间距	3
聚合物(Poly)掩模	聚合物宽度或间距	2
	有效扩散区域的栅极重叠区域	2
	栅极的有效重叠区域	2
	场聚合物与有效扩散区域的间距	1
P-选择,N-选择	沟道选择区域(重叠区域)	3
	有效扩散选择区域(重叠区域)	2
	触点选择区域(重叠区域)	1

（续）

掩 模 类 型	特 征 尺 寸	大小（λ 的倍数）
聚合物的简易触点	触点大小	2×2
	触点的有效重叠区域	2
	触点间距	2
	触点与栅极间距	2
聚合物的密集触点	触点大小	2×2
	触点的聚合物重叠区域	1
	相同聚合物上的触点间距	2
	不同聚合物上的触点间距	5
	与触点聚合物的接触	4
	与有效短沟道元器件的间距	2
	与有效长沟道元器件的间距	3
有效扩散区域的简易触点	触点大小	2×2
	触点的有效重叠区域	2
	触点间距	2
	触点与栅极的间距	2
有效扩散区域的密集触点	触点大小	2×2
	触点的有效重叠区域	1
	相同有效扩散区域上的触点间距	2
	不同有效扩散区域上的触点间距	6
	不同有效扩散区域的触点	5
	触点与栅极的间距	2
	场聚合物短沟道器件的触点	2
	场聚合物长沟道器件的触点	3
金属 1	宽度	3
	金属 1 之间的间距	3
	触点和聚合物的重叠区域	1
	触点和有效扩散区域的重叠区域	1
通道	大小	2×2
	通道之间的隔离	2
	金属 1 和通道的重叠区域	1
	与聚合物或有效扩散边缘的距离	2
	通道与触点的间距	2
金属 2	宽度	3
	金属 2 之间的间距	4
	通道的金属重叠区域	1
玻璃罩	焊接区（带有金属 2 底部沟槽）	100×100μm
	探针焊点	75×75μm
	焊点与玻璃罩边缘的距离	6μm

图 6-3 给出了带有功耗负载的 NMOS 反相器以及 CMOS 的规划示例，其中包含了上面讨论过的 NMOS 和 CMOS 可升级设计规则。图 6-3a 和图 6.3b 分别是 NMOS 和 CMOS 的逻辑电路图。在这两个图中，参数 Z 是指晶体管栅极的长度与宽度的比值。图 6-3c 和图 6-3d 分别给出了 NMOS 和 CMOS 的电路布置图。

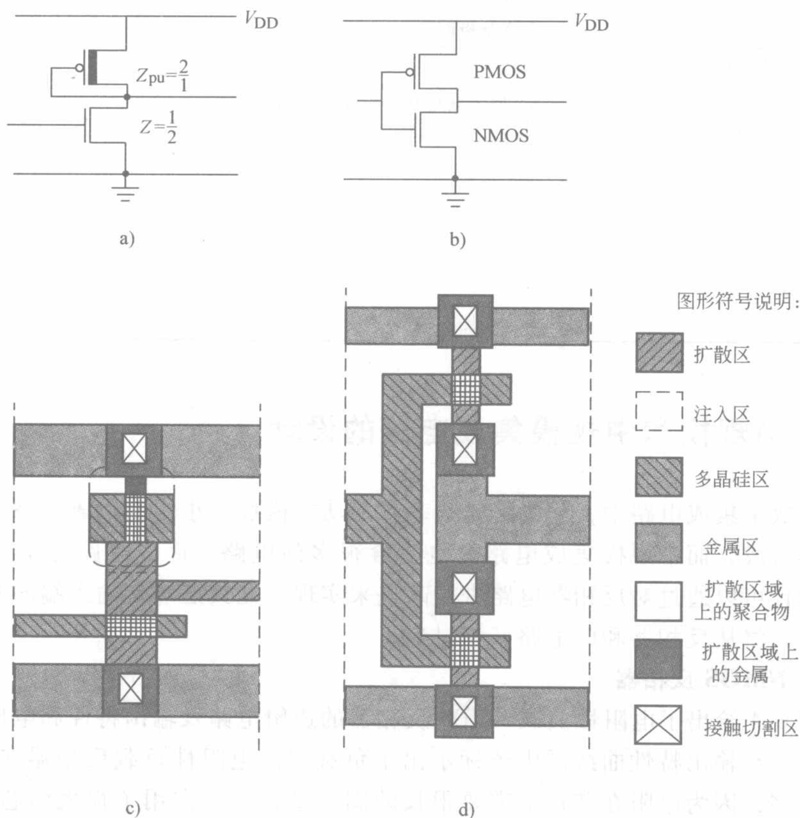


图 6-3　带有功耗负载的 NMOS 反相器以及 CMOS 的电路布置图设计示例
a）NMOS 反相器　b）CMOS 反相器　c）NMOS 反相器电路布置图
d）CMOS 反相器电路布置图

相对前面讨论过的 NMOS 和 CMOS，表 6-4 给出了双极性 NPN 晶体管的简化设计规则。如前所述，电路布置图中的最小特征尺寸由 λ 表示。表 6-4 中的 6 个掩模是制造工艺中必不可少的，包括：N^+ 型掩埋层扩散掩模、P^+ 型掩埋层扩散掩模、P 型基极扩散掩模、N^+ 型发射极和集电极扩散掩模、接触窗口掩模以及电路排版掩模。

表 6-4 NPN 晶体管的简化设计规则

掩 模 类 型	特 征 尺 寸	大小（λ 的倍数）
隔离掩模	隔离墙宽度	1
	隔离墙边缘至掩埋层的间距	2.5
基极掩模	基极至隔离区域的间距	2.5
发射极掩模	面积	2×2
	发射极与基极扩散区域的间距	1
	多重基极间距	—
集电极触点掩模	面积	1×5
	N^+ 型区域与基极扩散区域的间距	1
接触窗口掩模	基极触点	1×2
	发射极触点	1×1
	集电极触点	1×2
	基极触点与发射极的间距	1
电镀掩模	宽度	1.5
	金属间距	1

6.4 小规模和中规模集成电路的设计

在数字集成电路中，门电路是最基本的功能模块。小规模集成电路基本上由门电路组成，而中规模集成电路中也包含很多门电路。而门电路是基于反相器的，而且可以通过对反相器电路进行改进来实现，尤其是对多输入端的改进。本节接下来将从反相器和门电路开始讨论。

1. NMOS 反相器

图 6-4 给出了电阻性负载 NMOS 反相器的逻辑电路及输出特性和电压转移特性曲线，在输出特性曲线图中还标示出了负载线。电阻性负载反相器在 IC 中使用的不多，因为电阻在芯片上需要很长的固定条，从而占用了很大的芯片面积。解决这个问题的方法是采用有源负载，因为晶体管占用的芯片面积相对来说比较小。

图 6-5 给出了带有 3 种类型 NMOS 有源负载的 NMOS 反相器：饱和增强型负载、线性增强型负载和功耗负载。这些反相器之间进行比较的基础是几何比率——K_R。K_R 由 Z_{pu} 和 Z_{pd} 定义；其中，参数 Z 是指晶体管沟道长度与宽度的比值；下标 pu 是指上拉器件或负载器件；下标 pd 是指下拉晶体管或驱动晶体管。

饱和增强型负载反相器可以克服电阻性负载反相器在面积上的劣势。但是，当运行相同电流并具有相同下拉晶体管作为阻抗性反相器时，饱和增强型负载反

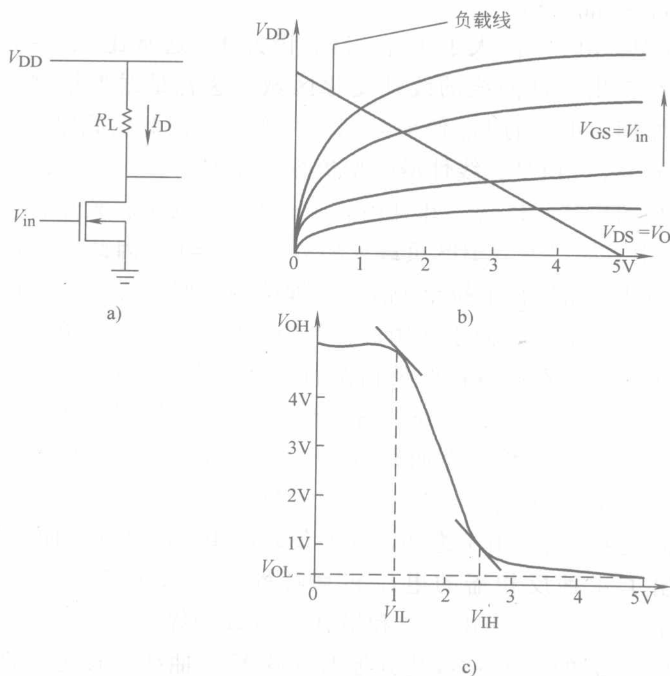


图 6-4　NMOS 电阻性负载反相器

a）电阻性负载 NMOS 反相器　b）输出特性曲线　c）转移特性曲线

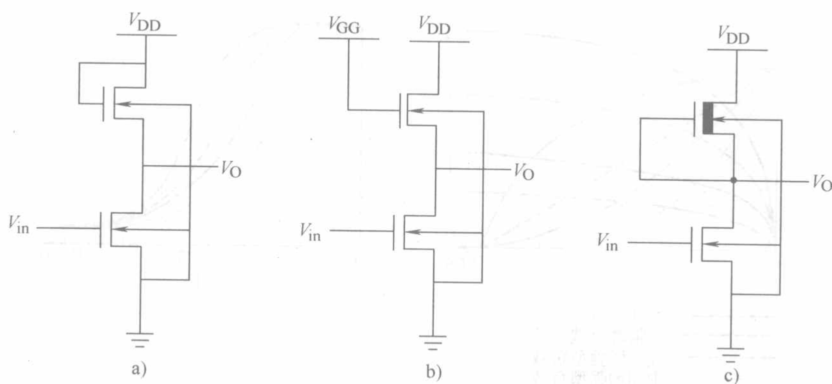


图 6-5　不同类型的有源负载 NMOS 反相器

a）饱和增强型负载　b）线性增强型负载　c）功耗负载

相器的 K_R 值就会很大，这表示负载电阻的面积还可能变得更小。但是，这种结构导致了与电阻性负载相关的微小逻辑电压摆动的产生，因为当负载晶体管的 $V_{GS}=V_{DS}$ 下降到 V_T 时，其工作状态就处于截止状态了。因此，对于这种类型的

反相器，其 $V_{OH}=V_{DD}-V_T$。

在图 6-5b 中，由于 V_{GG}大于 $V_{DD}+V_T$，因此 V_{DS}通常比 $V_{GS}-V_T$ 小；这样，负载通常运行在输出特性曲线的线性变化区域，这就是线性增强型负载 NMOS 反相器的工作原理。更大的 V_{GG}使得 V_{GS}比 V_T 更大，因此负载保持导通状态，而且 V_{OH}被上拉至 V_{DD}。但是，线性增强型负载反相器需要一个比饱和增强型负载反相器更大面积的负载晶体管，并且还需要为 V_{GG}触点添加额外的芯片面积。

在图 6-5c 中，耗尽型 NMOS 负载反相器的 $V_{GS}=0$，因此负载器件始终处于导通状态，而且 V_{OH}始终被上拉至 V_{DD}，这种结构克服了面积劣势而且不会产生电压摆动。因此，我们更倾向于使用这种类型的反相器。图 6-6a 和图 6-6b 给出了带有 4 种不同类型负载的 NMOS 反相器的性能比较。两条负载线和电压转移特性曲线都可以通过 SPICE 仿真得到。图 6-6a 中的负载线附加在下拉晶体管的输出转移特性曲线上，该转移特性曲线对于 4 种反相器来说是相同的。其中，R_L 为 100kΩ，每个反相器的 $V_{DD}=5V$，$V_{OL}=0.2V$，$I_{Dmax}=48\mu A$。我们可以注意到 V_{OH}达不到 V_{DD}，这主要是由于饱和增强型负载反相器的缘故，而不是其他原因。图 6-6b 还给出了 4 种反相器的电压转移特性（Voltage Transfer Characteristic，VTC）曲线。图中，V_{OH}仍然由于饱和增强型负载的缘故而低于 V_{DD}。我们还可以注意到，功耗负载的 VTC 曲线比其他器件的 VTC 曲线更接近于理想反相器的 VTC 曲线。

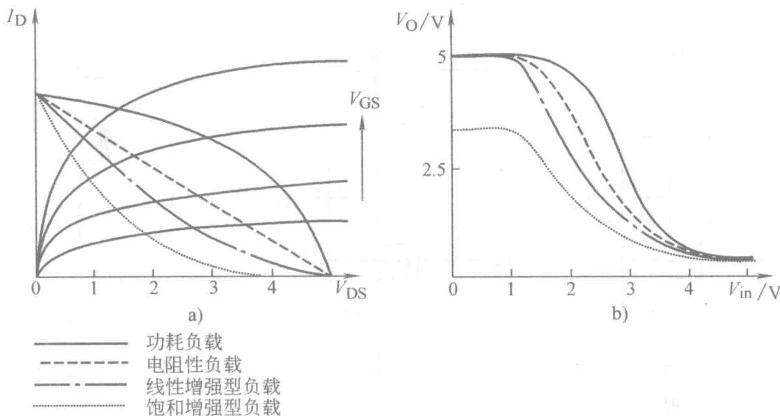


图 6-6 带有不同类型负载的 NMOS 反相器的性能特征曲线

a）输出特性曲线和负载线 b）电压转移特性曲线

图 6-6a 中的负载线很容易得到。例如，假设负载为耗尽型 NMOS。V_{GS}被固定在 0V，这样其输出转移特性曲线只包含了 $V_{GS}=0$ 对应的曲线。I_D 对于负载晶体管和驱动晶体管来说是相同的，但是它们的 V_{DS}被上拉至 V_{DD}。因此，两个晶

体管的 V_{DS}一个高，一个低。通过将负载特征曲线的 V_{DS}起点迁移到 V_{DD}就可以得到负载线了，这反映出了其纵轴随 V_{DD}变化以及和驱动反相器 V-I 曲线相重叠的特点。

电压转移特性曲线最好由计算机仿真产生。但是，在对关键电压（如 V_{OH}、V_{OL}、V_{IH}、V_{IL}以及与 V_{in}对应的 V_O）的分析中，我们仍然可以发现很多有用的信息。其中，NMOS 的电流是此类分析过程中的关键。增强型 NMOS 和耗尽型 NMOS 的开启电压是不同的，但是其漏极电流计算公式是相同的。线性区域和饱和区域的漏极电流计算公式分别如下：

$$I_D = K_n[2(V_{GS} - V_T)V_{DS} - V_{DS}^2];V_{DS} \leqslant V_{GS} - V_T \tag{6-3}$$

$$I_D = K_n(V_{GS} - V_T)^2;V_{DS} \geqslant V_{GS} - V_T \tag{6-4}$$

式中，V_T 为开启电压；$K_n = \frac{\mu_n C_{ox}}{2}\left(\frac{w}{L}\right)$为跨导；$\mu_n$为电子沟道迁移率；$C_{ox}$为单位面积的栅极电容。

上述定义的这些参数同样适用于 PMOS。

假设图 6-7a 为一个功耗负载 NMOS 反相器电路的 VTC 曲线图。在 $0 < V_{in} < V_T$的区域，驱动晶体管处于截止状态，因此，$V_{OH} = V_{DD}$。在 A 点处，V_{in}非常小。这样，对于驱动晶体管，其 $V_{DS} = V_O > V_{in} - V_T = V_{GS} - V_T$；对于负载，其 $V_{DS} = V_{DD} - V_O$。因此，驱动晶体管就处于饱和状态，负载就处于线性变化阶段。类似的分析同样适用于器件的其他工作阶段，如图 6-7 所示。对于两种晶体管来说，

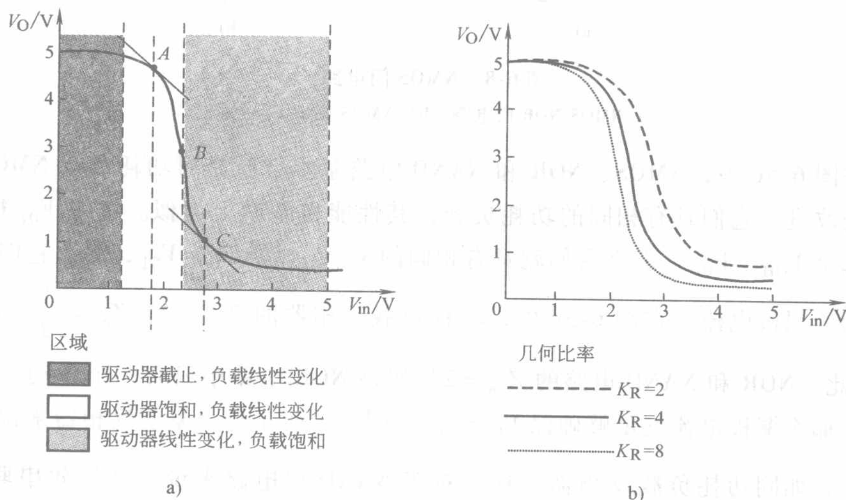


图 6-7　VTC 区域及功耗负载 NMOS 反相器不同几何比例的 VTC 曲线图

a）功耗负载 NMOS 反相器 VTC 曲线图及区域　b）不同几何比例的 VTC 曲线图

为了得到 V_{IH} 和 V_{IL}，A 和 C 点处的漏极电流计算公式是相同的。根据 V_{in} 来区分上述最后的公式，并根据条件：$dV_O/dV_{in}=-1$，就可以得到理想的关键电压。使饱和负载和线性驱动器的漏极电流在 $V_{in}=V_{DD}$ 条件下相等并对等式进行求解，就可以得到 V_{OL}。通过使任意 V_{in} 值处两个晶体管（晶体管工作在 V_{in} 对应的区域）的漏极电流相等，并根据给定的 V_{in} 值求解等式，就可以得到输出电压 V_O。

2. NMOS 门电路

NMOS 门电路中只需要考虑 NOR 和 NAND 电路，因为这两种电路在芯片面积上比 OR 和 AND 电路更经济，而且任何逻辑电路系统都可以由 NOR 或 NAND 电路来实现。通过将驱动晶体管并联来提供多个输入端，NMOS 反相器就可以很容易转换成 NOR 门电路，如图 6-8a 所示。而将驱动晶体管串联来，就可以得到 NAND 门电路。

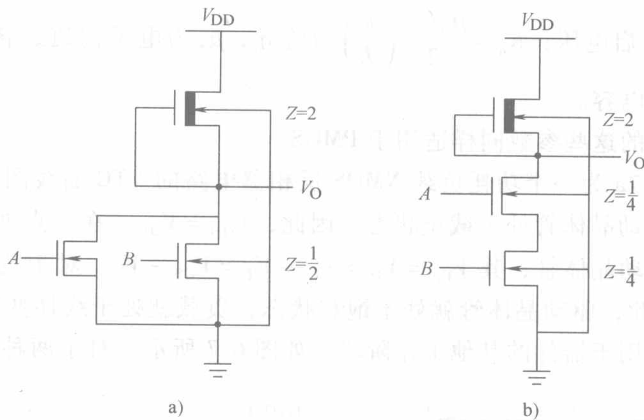


图 6-8　NMOS 门电路
a）NMOS NOR 门电路　b）NMOS NAND 门电路

在图 6-5c 中，NMOS、NOR 和 NAND 电路基本上就是对功耗负载 NMOS 反相器的改进。它们具有相同的功耗负载，其性能也基本上类似。如果 V_{DD} 相同，而且每个 $V_{OH}=V_{DD}$，那么它们就具有相同的 V_{OL}；如果 $V_O=V_{OL}$，那么它们就具有相同的漏极电流。在图 6-5c 中，功耗负载反相器的 $Z_{pu}=2$，$Z_{pd}=\frac{1}{2}$，$K_R=4$。因此，NOR 和 NAND 电路的 $Z_{pu}=2$。如果 NOR 电路中只有一个驱动反相器导通，那么漏极电流就足够确保 $V_O=V_{OL}$。因此，对于每个驱动反相器来说，其 $Z_1=\frac{1}{2}$，如同功耗负载反相器一样。对于 NAND 门电路来说，其等效串联 Z_{pd}（和漏源电阻对应）应该依然是$\frac{1}{2}$，从而 V_O 的值相同，而且 NAND 门电路中的

每个驱动晶体管的 $Z_I = \frac{1}{4}$。因此，反相器的 $K_R = 4$；NOR 门电路的 $K_R = 4$；NAND 门电路的 $K_R = 8$。随着输入端数量的增加，NAND 门电路的 K_R 也会增加，但是 NOR 门电路的 K_R 不会增加。因此，我们可以看出，相对 NMOS NOR 门电路来说，NMOS NAND 门电路更浪费芯片面积。从而，在 NMOS 电路中，我们更倾向于使用 NOR 门电路（而且 NOR 是标准的门电路）。

3. CMOS 反相器

如图 6-9a 所示，CMOS 反相器中包含一个增强型 NMOS 晶体管和一个互补增强型 PMOS 负载晶体管，并将增强型 NMOS 晶体管作为驱动晶体管。当 V_{in} 较小时，驱动晶体管处于截止状态，而当 V_{in} 较大时，负载晶体管处于截止状态。因此，这两个晶体管组成的串联电路通常是断路的，除非在切换时，两个晶体管都瞬间导通。低功耗是 CMOS 的一个重要优势，这个优势使其在 VLSI 设计中非常具有吸引力。

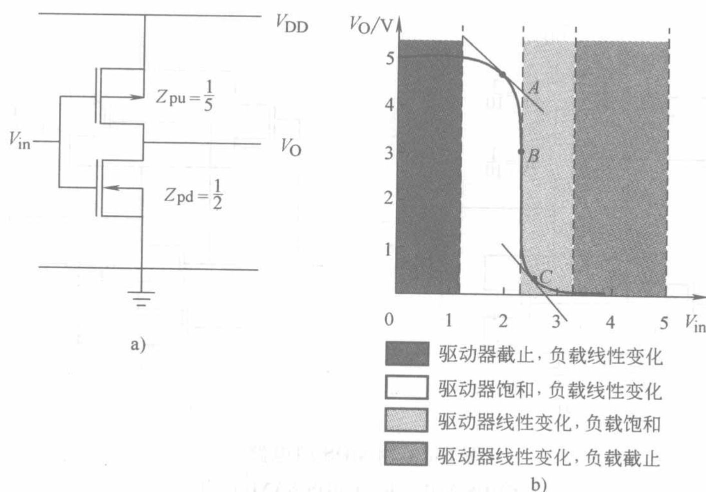


图 6-9　CMOS 反相器及其电压转移特性曲线

a）CMOS 反相器　b）CMOS 转移特性曲线及其工作区域

由于上拉器件从来没有断路过，而且 V_{OL} 由反相器比例决定，因此 NMOS 电路从这个意义上来说是一个比例电路；而 CMOS 电路则是一个无比例电路，因为其 V_{OL} 值始终是负值。但是，如果需要相同的源极电流和反向电流，那么上拉器件的电子迁移率/空穴迁移率的比值范围必须比下拉器件的大，典型值为 2.5 ~ 1。这使得电压转移特性曲线变成了对称曲线，其 $V_{in} = V_O$ 时的电压值为 $V_{DD}/2$，该电压称为“反相器电压 V_{inv}”。

图 6-9b 给出了 CMOS 反相器的电压转移特性曲线。我们可以注意到，该电

压转移特性曲线接近于理想逻辑反相器的电压转移特性曲线，这些特性都是通过计算机电路仿真程序得到的。如果是功耗负载 NMOS 反相器，我们还可以通过分析的方法得到更深入的认识。该分析过程同样适用其他类型的反相器。我们还可以注意到，图 6-9b 中的 VTC 曲线被划分成了各个区域，如同图 6-7a 一样。在每个区域中，负载和驱动晶体管的漏极电流都是相等的，以便在任何 V_{in} 下都可以计算出 V_O 的值。为了得到 V_{IL} 和 V_{IH}，在漏极电流计算公式中必须使用以下前提条件：$dV_O/dV_{in} = -1$。值得注意的是，PMOS 和 NMOS 的漏极电流计算公式是相同的（见式（6-3）和式（6-4）），除非 PMOS 的电压极性相反。

4. CMOS 门电路

CMOS 门电路是基于 CMOS 反相器的简单改进，图 6-10a 和 6-10b 给出了 CMOS NOR 和 NAND 门的电路图，其中 NOR 和 NAND 门电路基本上由 CMOS 反相器构成。在 CMOS 反相器中，负载和驱动晶体管分别由串联或并联（一定比例）的 PMOS 晶体管和 NMOS 晶体管代替。

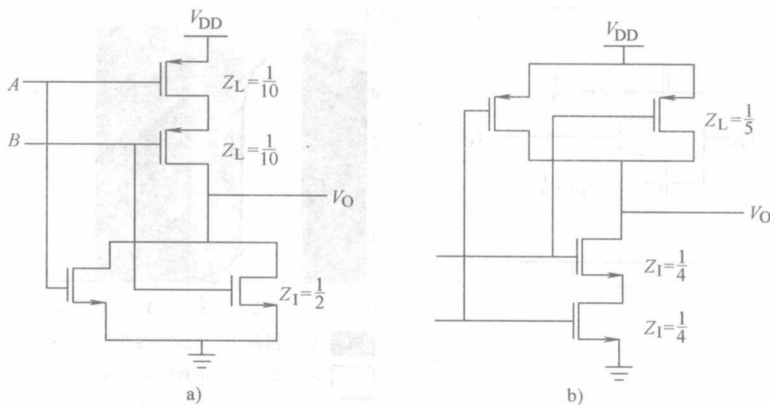


图 6-10　CMOS 门电路

a）CMOS NOR　b）CMOS NAND 门电路

假设图 6-10a 中 NOR 门电路的 V_{DD} 和 V_{inv} 与图 6-9a 中 CMOS 反相器的 V_{DD} 和 V_{inv} 相同，那么 NOR 门电路的 Z_{pu} 和 Z_{pd} 也应该与 CMOS 反相器的相同。在并联下拉晶体管中只需要一个导通来确保 $V_O = 0V$，$Z_I = Z_{pd} = \frac{1}{2}$，如同 CMOS 反相器的一样。但是，对于串联负载来说，$Z_L = \frac{1}{10}$ 等价于 $Z_{pu} = \frac{1}{5}$。如果图 6-10b 中的 NAND 门电路具有和上述反相器相同的 V_{inv}，那么其 $Z_I = \frac{1}{4}$，$Z_L = \frac{1}{5}$。这样，反相器的 $K_R = 0.4$；NOR 门电路的 $K_R = 0.2$；NAND 门电路的 $K_R = 0.8$（更接近整数）。因此，NAND 是 CMOS 电路中标准的门电路。另一种说法也可以表达上述

原理，即在给定的 Z 值下，如果沟道长度 L 不变，那么反相器、NOR 和 NAND 的负载宽度的比例为 1:2:1。

5. 双极性门电路

主要的双极性数字逻辑系列包括 TTL、发射极耦合逻辑（Emitter Couple Logic，ECL）元器件和集成注入逻辑（Integrated Injection Logic，I^2L）元器件。每个逻辑系列都可以进一步分类，例如，肖特基晶体管逻辑（Schottky Transistor Logic，STL）元器件和集成肖特基逻辑（Integrated Schottky Logic，ISL）元器件就是从基本 I^2L 系列发展而来的。双极性门电路具有比 CMOS 门电路更快的转换速度，但同时也具有更大的功耗。在 SSI 系列中最常用的双极性门电路是肖特基 TTL，该器件具有适中的功耗和传输延时。通过使用 I^2L 系列及其具有低功耗和适中转换速度的相关元器件，元器件的封装密度可以达到最高。这些逻辑系列之间的比较必须基于功率-延时乘积，该功率-延时乘积是指功耗与传输延时的乘积。

6. 中规模集成电路

中规模集成（MSI）电路的每个芯片上都包含了 10 ~ 100 个晶体管，这些晶体管都由反相器和基本逻辑开关构成，而且没有经过任何改进。它们都是经过最小化设计的，而不是将逻辑开关简单地堆积连接在一起。常见的 MSI 电路包括：触发器、计数器、寄存器、加法器、多路信号复用器以及多路信号分离器等等。

6.5 LSI 和 VLSI 电路设计

在 LSI 和 VLSI 设计中，半用户定制设计是一项使用很频繁的技术。在这项技术中，半成品子电路或单元电路通过互连形成更大的理想成品电路。这些子电路通常都是非常规范的，以便半用户定制设计技术可以得到非常规范的电路和规划布局。

多相时钟

多相时钟是一项非常重要的技术，它可以用来减少 LSI 和 VLSI 电路中元器件的数量。为了解释这种元器件数量上的节省原理，我们可以将对采用该技术的元器件数量与采用常规设计技术的元器件数量进行比较。在常规设计中，电路中包含了 1 个使用 D 形触发器（基于 CMOS NAND）的 4-b 移位寄存器和 1 个使用两相时钟和 CMOS 技术的 4-b 移位寄存器。

图 6-11 给出了这两种设计的逻辑图。图 6-11a 是移位寄存器的常规设计，该移位寄存器采用了单相时钟信号，而图 6-11b 给出了 D 形触发器的电路实现方法（Taub and Schilling，1977），该触发器使用了 CMOS NAND 门电路；在该设计中，元器件的数量如下所示：

1）5 个双输入 CMOS NAND 门电路，每个门电路包含 4 个晶体管，共计晶

体管数量为 20 个；

2）1 个三输入 CMOS NAND 门电路包含的晶体管数量为 6 个；

3）每个 D 形触发器包含的晶体管数量为 26 个；

4）4-b 移位寄存器包含的总晶体管数量为 104 个。

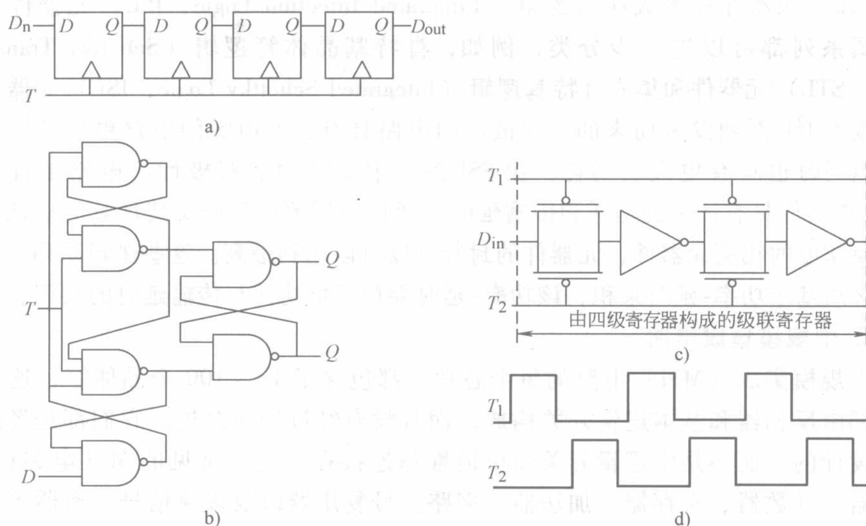


图 6-11 常规（静态）和动态 4-b 移位寄存器

a）常规静态移位寄存器 b）D 形触发器 c）动态移位寄存器 d）两相时钟脉冲

第二种设计中采用了二相时钟，如图 6-11c 所示，其互不重叠的时钟信号如图 6-11d 所示。我们可以注意到，每个触发器包含 2 个 CMOS 传输开关和 2 个 CMOS 反相器。因此，在 4-b 移位寄存器中包含了 8 个 CMOS 传输开关和 8 个 CMOS 反相器。在该设计中，元器件的数量如下所示：

1）8 个 CMOS 传输开关中的晶体管数量为 16 个；

2）8 个 CMOS 反相器中的晶体管数量为 16 个；

3）4-b 移位寄存器包含的总晶体管数量为 32 个。

从上面的例子中，我们可以发现，采用二相时钟最多可以将常规设计中的元器件数量减小至原来的 1/3。但是，这种优势在时钟和移位寄存器是动态的复杂情况下会被抵消掉。为了避免由于截止晶体管的渗漏而导致数据丢失，时钟必须大于一个最小频率，该最小时钟频率值由电容性负载的充放电时间决定。

6.6 MOS 电路中不断增加的封装密度和不断减小的功耗

CMOS 门电路的功耗比 NMOS 电路小，这在 VSL 和 VLSI 设计中是一个很大

的优势。但是，标准 CMOS 开关的每个输入端都需要两个晶体管，因此，其器件数量就比 NMOS 开关多，这是因为 NMOS 开关的每个输入端只需要一个晶体管，再加上一个功耗负载晶体管，该功耗负载晶体管与输入端无关（Mavor，Jack，and Denyer，1983）。NMOS 开关的这种优势被应用到了半导体存储器和可编程逻辑阵列中，半导体存储器和可编程逻辑阵列将在稍后讨论。除了需要更多的元器件外，还必须隔离 CMOS 中的 PMOS 晶体管和 NMOS 晶体管，并在它们的漏极之间采用金属导线进行互连，这两个漏极的电导率正好相反。因此，采用相同设计规则时，每个 NMOS 芯片上门电路的数量是 CMOS 芯片上的一半。

图 6-12 给出了一个 CMOS 多米诺逻辑电路的示意图，其中时钟用于非常规 CMOS 电路，用来实现高密度集成和低功耗。当 T 较低时，Q_1 截止，因此无论输入端 A、B、C、D 和 E 的电压是多少，都不会存在接地的路径；而此时 Q_2 导通，因此寄生电容 C_1 会被充电至 V_{DD}。当 T 为高电压时，Q_2 截止，Q_1 导通。因此，如果 A 和 B，或者 C 和 D，或者 A、B、C、D 同时为高电压时，那么从 C_1 至接地点之间的路径就到导通了，而且 C_1 将会放电。否则，C_1 上将会保持高电压（但电荷会慢慢流失），而且在输出端 F 处就会产生有效逻辑 $(AB)+(C+D)$。值得注意的是，该电路只包含两个负载 PMOS 晶体管，而且每个额外的输入端只需要一个驱动晶体管。因此，通过使用复杂逻辑功能代替简单逻辑功能可以使元器件数量最小化。除了输出端的反相器外，每个晶体管都可以最小化，因

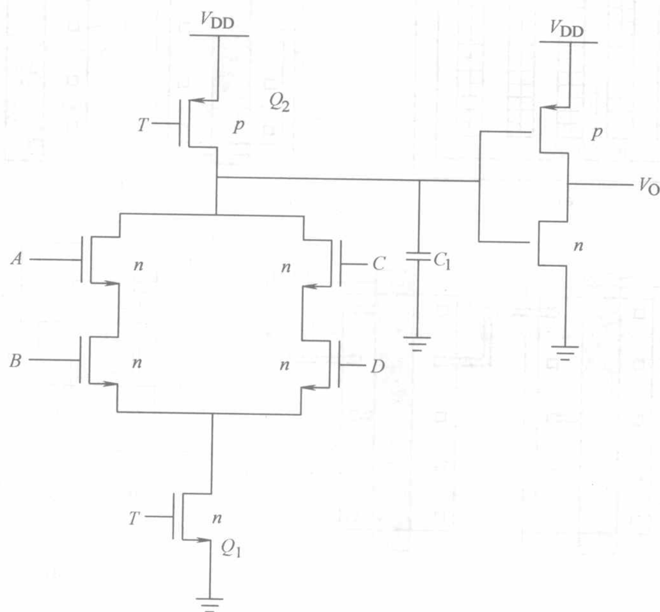


图 6-12　CMOS 多米诺 AND-OR 逻辑电路

为它们只需要用来对 C_1 进行充放电。由于不存在稳态电流，因此功耗如同标准CMOS一样低。

6.7 门阵列

门阵列是一种半用户定制的集成电路，其每个芯片上包含100多个至数千个纵横排列的开关单元。该开关单元可能是NAND或NOR电路，也可能是其他门电路。通常，每个开关单元都是一组元器件的集合体，这些元器件可以互连在一起以形成我们所需的门电路。各个开关单元的模式都是相同的，与芯片的功能无关。因此，门阵列可以预先进行大批量生产（Reinhard，1987）。门阵列的设计无需花费太多功夫，因为只有互连部分的掩模需要定制一个芯片，来满足特定用途的需要。

图6-13分别描述了不同详细程度的门阵列和开关单元互连部分的原理图。其中，图6-13a给出了门阵列的平面图，每个门阵列中包含了10个纵列，而每个纵列包含10个单元，因此，在该芯片上有100个开关单元。图6-13b给出了

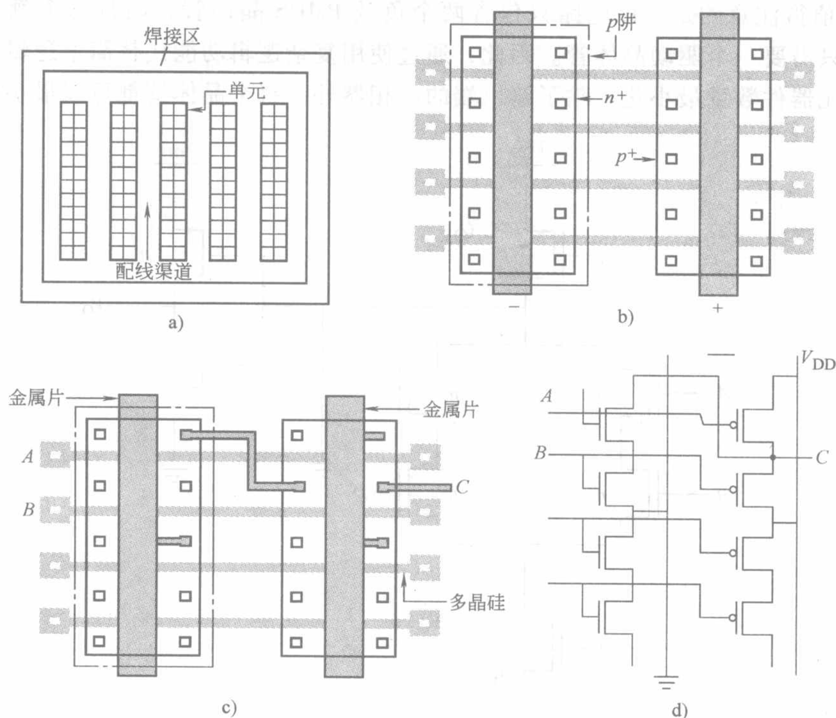


图6-13 各种不同详细程度的门阵列和开关单元互连部分原理图

a）开关单元结构 b）晶体管结构 c）NAND开关互连 d）等效电路

每个开关单元的结构图，其中每个开关单元中包含 4 个 NMOS 晶体管和 4 个 PMOS 晶体管。因此，在每个芯片上都包含 800 个晶体管。晶体管的沟道位于多晶硅下面的扩散区域中。图 6-13c 给出了开关单元形成 NAND 开关时的互连结构图，图 6-13d 给出了一个开关单元的等效逻辑电路图。

由于它们结构简单，因此在门阵列中就需要大量的配线来实现互连。性能良好的计算机软件可以用来实现互连部分的设计。而在实际中，配线渠道通常被填得很满，因此每个芯片上的开关单元的利用率很难超过 70%（Alexander，1985），接下来将要讨论的标准单元从某种程度上可以通过使用更复杂的逻辑功能或单元来缓解这个问题。

6.8 标准单元

在使用标准单元时，IC 设计师会从前面定义的逻辑电路或单元库中选择元器件来构建所需的电路。除了基本门电路外，单元库通常还包括更加复杂的逻辑电路，如单 OR、AND-OR-INVERT、触发器、加法器、只读存储器（ROM）等等。

标准单元设计方法非常适合自动制版，这个设计过程包括：根据所需电路的功能从单元库中选择逻辑单元、确定逻辑单元的相应位置，最后对它们进行互连。采用此方法设计的芯片平面图类似于图 6-13a 中门阵列芯片的平面图。但是，值得注意的是，在这种情况下，设计师可以控制配线渠道的数量和宽度。每个单元的布置图都是一样的，但是在使用时，在芯片中每个单元及其相关位置是惟一的。因此，该设计方法的每个掩模类型都是惟一的，而且其工艺比门阵列的设计方法更复杂，也更昂贵（Hodges and Jackson，1988）。

6.9 可编程逻辑器件

可编程逻辑器件（Programmable Logic Devices，PLD）已经广泛应用于 LSI 和 VLSI 的设计中，是用来实现两级电平、乘积和以及布尔功能的一类电路。多级逻辑器件可以通过 Weinberger 阵列或门矩阵来实现（Fabricius，1990；Weinberger，1967）。PLD 包括可编程逻辑阵列（Programmable Logic Arrays，PLA）、可编程阵列逻辑（Programmable Arrays Logic，PAL）以及 ROM。AND-OR 结构的 PLA 是所有 PLD 的核心，可用来实现任何两级电平。AND-OR 功能常由 NOR-NOR 或 NAND-NAND 实现。

PLD 的最大优势是可构建超常规的平面结构。PLD 由一个 AND 平面组成，其后面是一个 OR 平面。在每个单独的导电层中，逻辑功能由纵横交叉处触点或

连接点的连通情况来决定。换句话说，PLD的可编程功能是通过这些交叉点的熔合连接来实现的。

图6-14给出了3种类型的PLD。在AND平面和OR平面上，纵横交叉处的空心点表示该平面是可编程的。而纵横交叉处的实心点则表示该平面已经被定义或固定。如果AND平面和OR平面都是可编程逻辑平面，那么该PLD就是一个PLA逻辑器件；如果只有AND平面是可编程逻辑平面，那么该PLD就是一个PAL逻辑器件；如果只有OR平面是可编程逻辑平面，那么该PLD就是一个ROM逻辑器件（本例中为解码器）。由于PLA逻辑器件的两个平面都是可编程

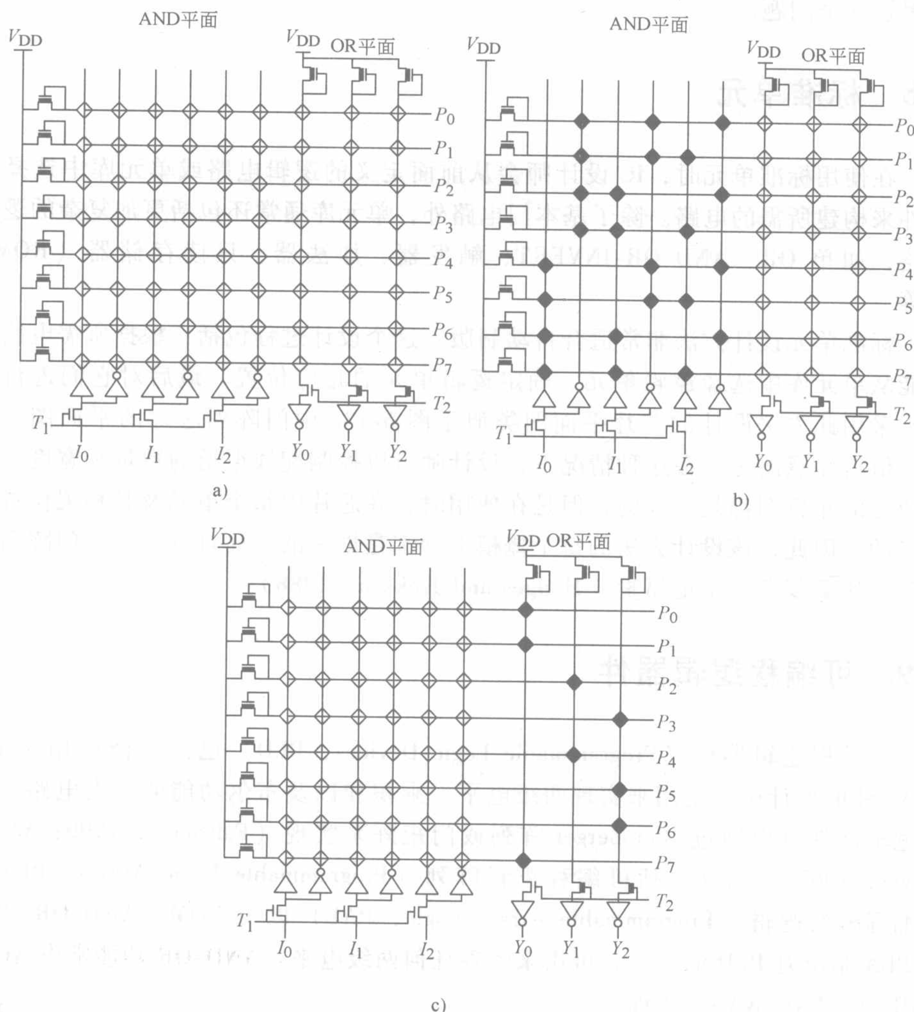


图6-14　各种类型的可编程逻辑器件

a）可编程逻辑阵列（PLA）　b）可编程只读存储器（ROM）　c）可编程阵列逻辑（PAL）

逻辑平面，这使得 PLA 在实现各种逻辑功能时比 PAL 更加灵活。这样，PAL 逻辑器件就可以看做是 PLA 的一种特殊情况。因此，接下来我们只讨论 PLA 逻辑器件。

可编程逻辑阵列

PLA 为各种用来构建超常规平面结构的组合逻辑功能实现提供了很好的选择。例如，假设一个 PLA 的乘积和如式（6-5）所示：

$$Y_1 = \bar{I}_0 \bar{I}_1 + I_0 \bar{I}_2 \tag{6-5}$$

$$Y_2 = \bar{I}_0 I_1 I_2 + I_0 \bar{I}_2 \tag{6-6}$$

$$Y_3 = \bar{I}_0 \bar{I}_1 \bar{I}_2 + \bar{I}_0 I_1 I_2 \tag{6-7}$$

PLA 具有 3 个输入端和 2 个输出端。以 AND 平面和 OR 平面为例，AND 平面的输出为

$$P_1 = \bar{I}_0 \bar{I}_2 \tag{6-8}$$

$$P_2 = \bar{I}_0 \bar{I}_1 \bar{I}_2 \tag{6-9}$$

$$P_3 = \bar{I}_0 I_1 I_2 \tag{6-10}$$

$$P_4 = I_0 \bar{I}_2 \tag{6-11}$$

电路的整体输出为 OR 平面的输出，并可以根据 AND 平面的输出得到

$$Y_1 = P_1 + P_4 \tag{6-12}$$

$$Y_2 = P_3 + P_4 \tag{6-13}$$

$$Y_3 = P_2 + P_3 \tag{6-14}$$

图 6-15 给出了由 AND 和 OR 平面构成的逻辑电路。值得注意的是，在 AND 平面上的每条乘积线都是一个 NMOS NOR 门电路，该 NMOS NOR 门电路带有一个功耗负载；而且每个驱动晶体管的门电路由输入线控制。同样，在 OR 平面上的每条输出线都是一个 NMOS NOR 门电路，该 NMOS NOR 门电路带有一个驱动晶体管，驱动晶体管的门电路由乘积线控制。因此，PLA 是由 NOR-NOR 实现的。

PLA 的特征矩阵（Lighthart, Aarts, and Beenker, 1986）可以用来很好地描述 PLA 的编程原理。式（6-15）给出了图 6-15 中 PLA 的特征矩阵

$$Q=\begin{bmatrix}0&0&x&1&0&0\\0&0&0&0&0&1\\0&1&1&0&1&1\\1&x&0&1&1&0\end{bmatrix}\tag{6-15}$$

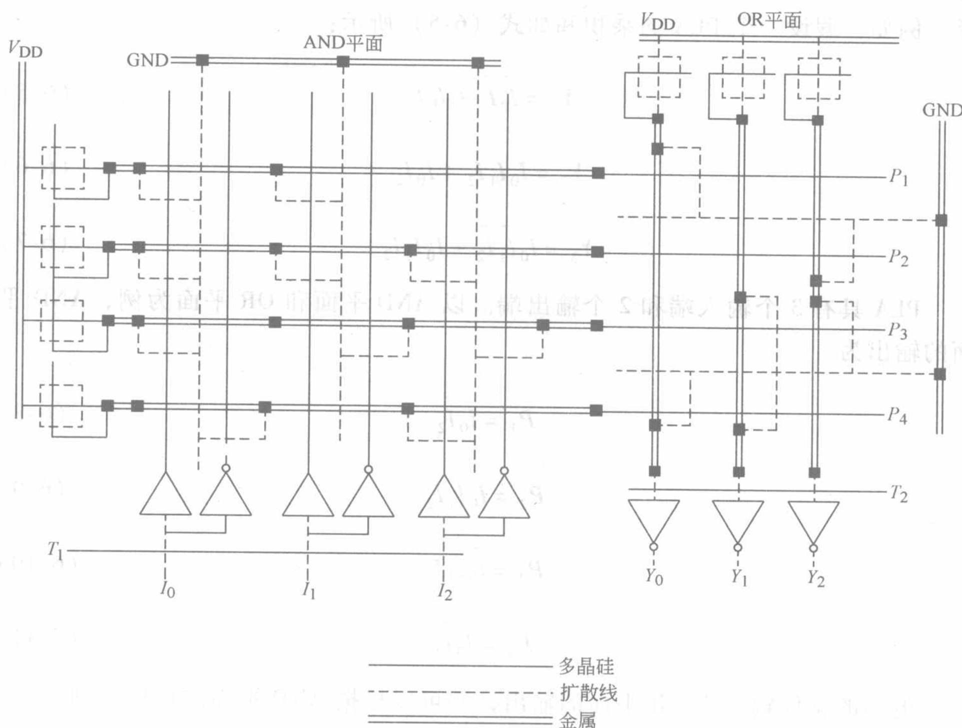


图 6-15 NMOS PLA 电路

前 3 列构成了 AND 平面的矩阵，后 3 列构成 OR 平面的 3 输入、3 输出 PLA。在 AND 平面上，如果晶体管用来将乘积线 P_i连接到输入线 I_i，那么矩阵元素 $q_{ij}=0$；如果晶体管用来将乘积线 P_i连接到输入线 $\bar{I}_i$，那么矩阵元素 $q_{ij}=1$，而且如果没有任何输入线连接到乘积线 P_i，那么矩阵元素 q_{ij}就变得无关紧要了。在 OR 平面上，如果乘积线 P_i连接到输出线 Y_j，那么矩阵元素 $q_{ij}=1$；其他情况下，$q_{ij}=0$。

图 6-16 给出了 PLA 的电路符号平面图，并解释了如何由 PLA 的常规结构得到它的平面图。连接到每个平面的输入线都是多晶硅，从各平面引出的输出线都是金属线，各驱动晶体管的源极都被连接扩散线接地。各晶体管通过接地的、互相交叉的扩散线相互连接。

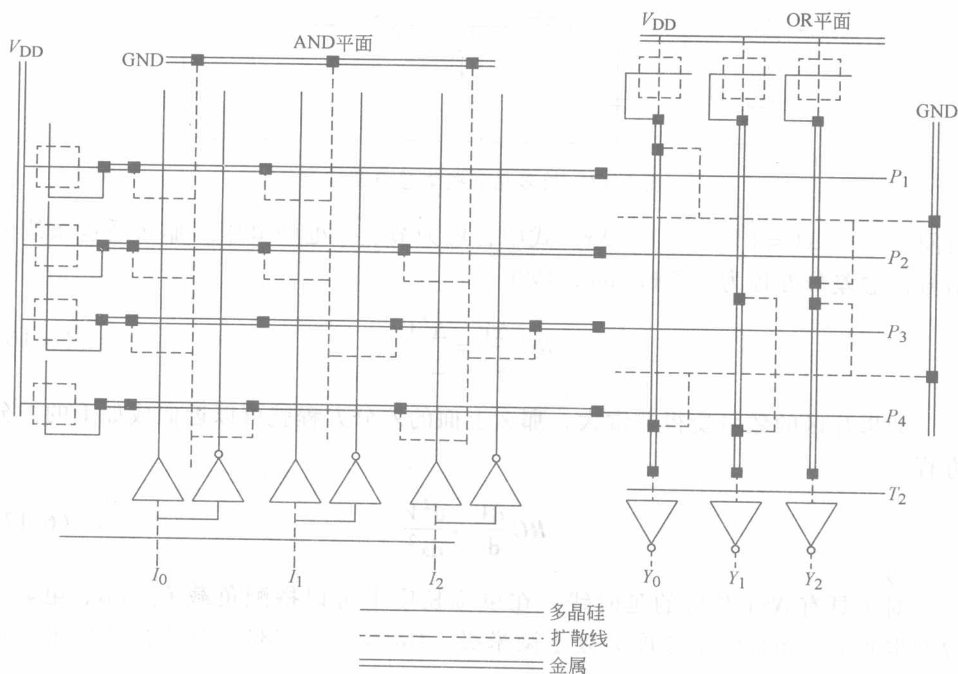


图 6-16　PLA 电路条形平面图

6.10　不断减小的传输延时

在大规模集成电路中，我们经常可以见到很大的电容性负载。连接焊盘用来将芯片与其他电路连接，而在测试时经常使用探测焊盘。这两者（焊接区和探测焊盘）都可以将电容性负载与其驱动器连接起来。芯片上的互连部分是通过金属线或多晶硅线连接起来的。当距离较长时，该金属线或多晶硅线可以将较长的电容性负载与其驱动器连接起来。尽管常规阵列结构（如门阵列、标准单元和 PLA）在 LSI 和 VLSI 半定制用户设计中使用很方便，但是它在传输延时方面存在先天性的缺陷。它们的纵横导线连接了很多元器件，因此，具有很高的电容性。在一条长连接线中，可以通过在沿线插入缓冲器来存储信号，从而减小整体延时。超级缓冲器通常用于连接芯片内部的小型门电路和大型焊盘驱动器，也可以用来驱动高电容性导线。

1. 电阻-电容（RC）延时线

较长的多晶硅线可以通过集总 RC 传输线来建模，如图 6-17 所示。假设 Δx 表示由电阻 R 和电容 C 构成的某个片区的长度，Δt 表示信号通过该片区所需的

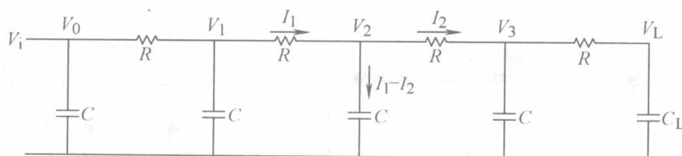


图 6-17 长多晶硅线集总电路模型

时间，而且 $\Delta V=(V_{n-1}-V_n)\Delta x$。式中，$V_n$为节点 n 处的电压。那么沿这条线的信号传输差分方程为（Fabricins，1990）

$$RC\frac{\Delta V}{\Delta t}=\frac{\Delta^2 V}{\Delta x^2} \tag{6-16}$$

如果片区的数量变得非常大，那么上面的差分方程就可以近似成如下的微分方程：

$$RC\frac{\mathrm{d}V}{\mathrm{d}t}=\frac{\mathrm{d}^2 V}{\mathrm{d}x^2} \tag{6-17}$$

对于具有 N 个片区的延时线，在单位长度上可以搭配负载 $C_L=C$、电阻 R 以及电容 C，而且整个长度 L 处于微米级，Horowitz（1983）为我们给出了传输延时的公式

$$t_d=\frac{0.7N(n+1)L^2RC}{2N^2} \tag{6-18}$$

从上面的公式我们可以看出，传输延时与多晶硅线长度的平方成正比。当 N 趋向于无穷大时，传输延时的公式就变成了

$$t_d=\frac{0.7RCL^2}{2} \tag{6-19}$$

为了减小整体延时，可以沿多晶硅线插入复原反相器。例如，假设一条多晶硅线的长度为 5mm，其 $r=20\Omega/\mu m$，$C=0.2fF/\mu m$。如果插入到多晶硅线的反相器的数量从 0 个变成 4 个，就必须计算出相应的传输延时。每个反相器的延时与其驱动的片区长度成正比，每驱动 1mm 长的片区，就会产生 $t_l=0.4ns$ 的延时。每个片区中，反相器都是沿多晶硅线均匀布置的。其参数如下所示：k 为使用的反相器数量；$k+1$ 为片区的数量；l 为每个片区的长度，$l=\frac{5mm}{k+1}$；t_d为整体延时。

因此，t_d可以计算如下：

$$t_d=\left[(k+1)\left(\frac{0.7RCl^2}{2}\right)+0.4lK\right]ns \tag{6-20}$$

从式（6-20）可以计算出传输延时，还可以计算出作为无缓冲线延时百分比的每一批反相器的延时。表 6-5 列出了计算结果。

表 6-5　随着缓冲线的增加而改进的传输延时

反相器 k 的序号	整体延迟 t_d/ns	无缓冲延迟百分比
0（无缓冲）	35.0	100
1	18.5	52.86
2	13.0	37.14
3	10.25	29.29
4	8.6	24.57

表 6-5 中的数据显示，随着反相器数量的增加，传输延时会逐渐减小。但是，传输延时的平均减小程度没有单个反相器的减小程度大。因此，当增加的增益无法再调整添加的反相器时，设计师可能会停止增加反相器的数量。如果反相器的数量是偶数，那么整体信号就不会产生反相。

2. 超级缓冲器

超级缓冲器可以用来减小传输延时，而且不会产生过多的功耗。这些超级缓冲器都是反相或非反相电路，可以用来产生或衰减比标准反相器更大的电流，并驱动比标准反相器更快的大型电容性负载。在 NMOS 反相器中，上拉电流驱动能力远小于下拉驱动能力；与正比例 NMOS 反相器不同，超级缓冲器具有均衡的驱动能力。超级缓冲器由一个推挽式或“图腾柱”输出反相器构成，该反相器由常规反相器驱动。在反相缓冲器中，下拉驱动晶体管和“图腾柱”反相器的门电路都是由输入信号驱动，而输出“图腾柱”反相器的上拉晶体管的门电路是由输入信号的补码来驱动的。

图 6-18 给出了反相和非反相 NMOS 超级缓冲器的逻辑电路图。如果将反相器比例（K_R）设为 4，并将“图腾柱”上拉晶体管的驱动电压设置成标准耗尽型上拉晶体管电压的两倍，那么 NMOS 缓冲器就基本上变成了无比例电路。在标准 NMOS 反相器中，上拉晶体管的转换速度相对慢一些。图 6-18a 即为反相缓冲器。当输入电压变小时，标准反相器的输出电压和 Q_4 的栅极电压会迅速变大，因为相关的负载只有较小的栅极电容 Q_4。这样，“图腾柱”电流的输出转换就很快了。以上的分析同样适用于非反相缓冲器中，最后也会产生很快的转换速度。

相比标准 NMOS 反相器，NMOS 缓冲器电流驱动能力的改善程度可以通过比较平均输出上拉电流来进行评估（Fabricius，1990）。标准 NMOS 反相器的功耗负载处于饱和状态时，$V_O < 2V$；处于线性变化区域时，$V_O > 2V$。对于上拉器件，$V_{DS} = 5V - V_O$。这样，当 $3V < V_{DS} < 5V$ 时，上拉晶体管就处于饱和状态；当

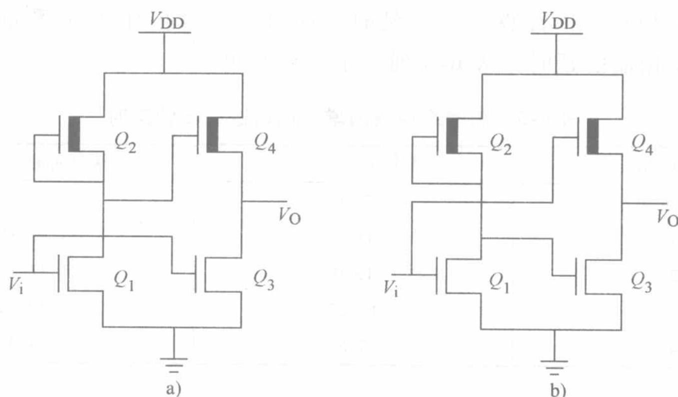


图 6-18 NMOS 缓冲器
a) 反相缓冲器 b) 非反相缓冲器

$0V < V_{DS} < 3V$ 时，上拉晶体管就处于线性变化区域。平均电流可以通过计算 $V_{DS}=5V$ 时的 $I_{D(sat)}$ 和 $V_{DS}=2.5I_{D(lin)}$ 来确定。假设耗尽型晶体管的 $V_{TD}=-3V$，标准 NMOS 反相器的电流计算公式如下：

$$I_{D(sat)}=K_{pu}(V_{GS}-V_{TD})^2=K_{pu}(0+3)^2=9K_{pu} \tag{6-21}$$

$$I_{D(lin)}=K_{pu}(2(V_{GS}-V_{TD})V_{DS}-V_{DS}^2)=K_{pu}(2(0+3)2.5-2.5^2)=8.75K_{pu} \tag{6-22}$$

因此，标准 NMOS 反相器的平均上拉电流大约为 $8.88K_{pu}$。对于 NMOS 缓冲器的“图腾柱”输出来说，平均上拉电流由 $V_{DS}=5V$ 和 2.5V 时的漏极电流来确定。值得注意的是，在这个例子中，上拉晶体管导通时，其 $V_G=V_{DD}=5V$。这样，$V_{GS}=V_{DS}=5V$，因此该晶体管始终处于线性工作区域。各电流如下：

$$I_D(5V)=K_{pu}[2(5+3)5-5^2]=55K_{pu} \tag{6-23}$$

$$I_D(2.5V)=K_{pu}[2(2.5+3)2.5-2.5^2]=21.25K_{pu} \tag{6-24}$$

“图腾柱”输出的平均上拉电流为 $38.12K_{pu}$。平均“图腾柱”上拉电流大约是平均 NMOS 上拉电流的 4.3 倍。因此，如果反相器的设计比例为 4，缓冲器基本上是无比例的。

6.11 输出缓冲器

一个 VLSI 芯片的内部开关具有大约小于 50fF 的负载电容以及小于 1ns 的典型传输延时。但是，芯片的输出端必须驱动 50pF 或更大的电容性负载（Hodges

and Jackson，1988）。对于 MOSFET 来说，传输延时直接与负载电容成正比。这样，在芯片上使用一个典型的门电路去驱动一个输出端将会造成传输延时过长。输出缓冲器通过级联具有很强驱动能力的反相器来减小传输延时。

图 6-19 给出了一个 N 级输出缓冲器的例子。这种很强的驱动能力来自于不断增加的晶体管沟道宽度。当晶体管的宽度因子 f 从一级增加到另一级时，电流驱动能力和输入电容也会逐级增加。如果 C_G 表示缓冲器中第一个反相器的输入或开关电容，那么第二个反相器的输入或开关电容就是 fC_G。依此类推，第 N 个反相器的输入或开关电容就是 $f^{N-1}C_G$，而负载电容就是 f^NC_G，该值等于输出端的负载电容 C_L。图 6-19 中左边的反相器是芯片上一个典型的反相器，其输入或开关电容为 C_G，传输延时为 τ。缓冲器中的第一个反相器的输入电容为 C_G，但是它的驱动能力比芯片上的反相器大 f 倍。因此，其传输延时为 $f\tau$。缓冲器中的第二个反相器的输入电容为 fC_G，其输出端的传输延时为 $2f\tau$。依此类推，第 N 个反相器的输入电容就是 f^NC_G，该值等于输出端的负载电容，其传输延时为 $Nf\tau$，该传输迟延就是缓冲器整体的延时。

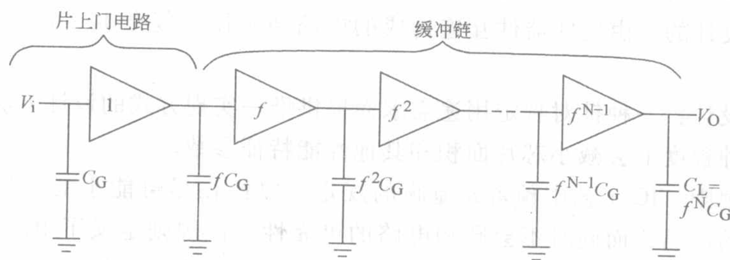


图 6-19　N 级输出缓冲器链

假设：

$$负载电容 = YC_G = f^NC_G$$
$$整体延时 = t_B = Nf\tau$$

那么，

$$N = \frac{\ln Y}{\ln f} \tag{6-25}$$

$$t_B = \frac{\ln Y}{\ln f} f\tau \tag{6-26}$$

将第一个引出的 t_B 置零，并根据 f，我们可以计算出，当 $f = e = 2.72$（即自然对数的基）时，t_B 最小。该最小值不是突发性的（Moshen and Mead，1979），而且 f 的值处于 2 和 5 之间时，不会再增加更多的延时。

例如，一个典型门电路的 $C_G = 50\text{fF}$，$\tau = 0.5\text{ns}$，它可以用来驱动芯片中相

同的门电路。但是，如果我们假设这个门电路用来驱动另一个输出端而不是驱动相同的门电路，该输出端的负载电容为 $C_L = 55\text{pF}$。如果使用输出缓冲，那么

$$Y = \frac{C_L}{C_G} = \frac{55\text{pF}}{50\text{fF}} = 1100$$

$$N = \ln Y = 7$$

$$t_B = 7e\tau = 9.5\text{ns}$$

如果将典型芯片的门电路直接连接到输出端，那么传输延时就变成了 $Y\tau = 550\text{ns}$，该延时比使用缓冲器时的 9.5ns 延时大了很多。因此，这个例子就很好地解释了缓冲器的作用。

名词解释

单元库：简单逻辑器件的集合体，这些逻辑器件是根据特定设计规则和制造工艺进行设计的。由这些器件互连构成的电路通常用于复杂 IC 芯片的半定制设计中。

定制设计：一种针对特定用途需求而提供惟一实现方式的设计方法，该特定用途在某种程度上会减小芯片面积和其他性能特征参数。

设计规则：IC 工艺中预备光掩膜的规定，以便在尽可能小的几何尺寸内得到最合适的产品，而同时不会影响电路的可靠性。该规则定义了 IC 芯片上最小的元器件尺寸、特征尺寸间最小的间隔以及最大的特征尺寸偏差。

电路布置或排版：IC 设计中一个重要的步骤，该步骤定义了不同掩模层上元器件的位置和特征尺寸大小。

掩膜：用来定义扩散区域、电路排版等区域的一组光学涂层，位于 IC 晶片层之上。每个掩膜由惟一的模式构成，即相应层的镜像。

标准单元：单元库中根据特定设计规则和制造工艺预先设计的逻辑电路。标准单元通常用于复杂电路的半定制设计中。

半定制设计：通过将预先设计好的子电路或单元连接起来，以形成所需的复杂电路或其中一部分电路的设计方法。

参考文献

[1] Alexander, B. 1985. MOS and CMOS arrays. In *Gate Arrays: Design Techniques and Application*. ed. J. W. Read. McGraw-Hill, New York.

[2] Fabricius, E. D. 1990. *Introduction to VLSI Design*. McGraw-Hill, New York.

[3] Hodges, A. D. and Jackson, H. G. 1988. *Analysis and Design of Digital Integrated Circuits*. McGraw-Hill,

New York.

[4] Horowitz, M. 1983. Timing models for MOS pass networks. *Proceedings of the IEEE Symposium on Circuits and Systems*, pp. 198-201.

[5] Keyes, R. W. 1975. Physical limits in digital electronics. *Proc. of IEEE* 63: 740-767.

[6] Lighthart, M. M., Aarts, E. H. L., and Beenker, F. P. M. 1986. Design for testability of PLAs using statistical cooling. *Proceedings of the 23rd ACM. IEEE Design Automation Conference*, pp. 339-345, June 29-July 2.

[7] Mavor, J., Jack, M. A., and Denyer, P. B. 1983. *Introduction to MOS LSI Design*. Addison-Wesley, Reading, MA.

[8] Mead, C. A. and Conway, L. A. 1980. *Introduction to VLSI Systems*. Addison-Wesley, Reading, MA.

[9] Moshen, A. M. and Mead, C. A. 1979. Delay time optimization for driving and sensing signals on high capacitance paths of VLSI systems. *IEEE J. of Solid State Circ.* SC-14 (2): 462-470.

[10] Pucknell, D. A. and Eshroghian, K. 1985. *Basi VLSI Design, Principles and Applications*. Prentice-Hall, Englewood Cliffs, NJ.

[11] Reinhard, D. K. 1987. *Introduction to Integrated Circuit Engineering*. Houghton-Mifflin, Boston, MA.

[12] Taub, H. and Schilling, D. 1977. *Digital Integrated Circuits*. McGraw-Hill, New York.

[13] USC Information Sciences Inst. 1984. MOSIS scalable NMOS process, version 1.0. Univ. of Southern California, Nov., Los Angeles, CA.

[14] USC Information Sciences Inst. 1988. MOSIS scalable and genertic CMOS design rules, revision 6. Univ of Southern California, Feb., Los Angeles, CA.

[15] Weinberger, A. 1967. Large scale integration of MOS complex logic: a layout method. *IEEE J. of Solid-State Circ.* SC-2 (4): 182-190.

备注

IEEE 发行的关于固态电路的刊物包括：

[1] 《IEEE 特定集成电路会议》学报。

[2] 《IEEE 电子器件》学报。

[3] 《欧洲固态电路会议（ESSCIRC）》学报。

[4] 《IEEE 设计自动化会议》学报。

第7章 数字逻辑系列

Robert J. Feugate, Jr.

7.1 引言

数字器件通常工作在两个稳定的工作区域（通常是指电压范围），这两个工作区域被过渡区域隔离开；而器件的工作状态只会穿过过渡区域，而不会停留在其中，如图7-1所示。如果器件穿过过渡区域的时间过长，虽然不会对器件产生不利影响，但是会导致输出不稳定。在图7-1中，当输入电压低于指定的 V_{IH} 时，反相器的输出电压肯定比 V_{OH} 大。值得注意的是，该电路在设计时，其输入高电压要低于输出高电压（对于逻辑低电压来说也是如此）。这种差异称为"噪声容限"或"噪声安全系数"，噪声容限允许干扰信号在一定程度上影响逻辑电压，但不会导致电路产生错误操作。逻辑条件和二进制数值之间的相互关系可以描述如下：其中一个稳态电压范围与逻辑状态或二进制数值相互关联，同时其他的稳态电压与相反的逻辑状态或二进制数值相互关联。这样，通过扩展，我们就可以进行电路设计并执行逻辑操作或算术运算。由于逻辑电路的详细设计不在本章涵盖的内容之中，读者可以参考其他关于数字逻辑电路设计的教材⊖。

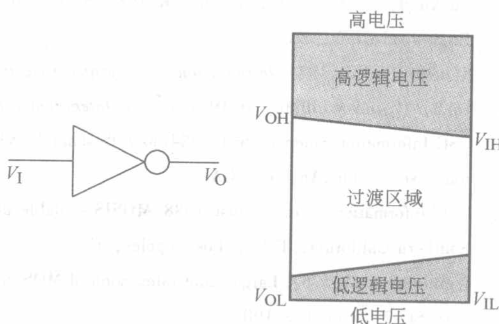


图7-1 逻辑电压等级

数字逻辑元器件是最早商业化生产的集成电路。其中，电阻-晶体管逻辑（Resistor-Transistor Logic，RTL）电路和速度更快的电阻-电容-晶体管逻辑（Resistor-Capacitor-Transistor Logic，RCTL）电路最早出现在20世纪60年代，由Fairchild半导体公司生产；而稍后出现的二极管-晶体管逻辑（Diode-Transistor Logic，DTC）电路由Signetics公司生产。尽管这些系列的逻辑元件在电子学书籍

⊖ 两本最新的关于逻辑电路设计的书：*Modern Digital Design*, by Richard Sandige, McGraw-Hill, New York, 1990 和 *Contemporary Logic Design*, by Rand Katz, Benjamin Cummings, Redwood City, CA, 1994.

中通常被介绍成后来逻辑系列的鼻祖，但是它们已经被弃用很多年了，只剩下一些历史意义而已了。

不同逻辑系列之间的主要性能差异在于它们的速度-功率乘积，即一个基本开关单元的平均传输延迟乘以其平均功耗。表7-1列出了几种流行逻辑系列的速度-功率乘积值。需要注意的是，制造商手册里面指定的传输延迟是在不同负载条件下测得的，这些负载条件在计算速度-功率乘积时必须考虑在内。

表7-1 离散逻辑系列①的速度-功率乘积之间的比较

逻辑系列	功耗门电路			
	静态噪声(100kHz/mW)	典型延迟/mW	门电路	乘积(100kHz/ns)
金属栅极 CMOS	0.001	0.1	75	7.5
硅栅极 CMOS	0.0000025	0.17	10	1.7
标准 TTL	10	10	10	100
低功率肖特基 TTL	2	2	8	16
肖特基 TTL	19	19	4.5	85.5
先进肖特基 TTL	8.5	8.5	1	8.5
ALS TTL	1	1	2.5	2.5
10K ECL	24	24	0.2	4.8

① 负载电容为50pF；负载电阻为500Ω，除非ECL正在驱动50Ω的传输负载。

7.2 晶体管-晶体管逻辑电路

晶体管-晶体管逻辑（TTL）电路最早出现在20世纪60年代，在90年代的离散信号逻辑电路设计中可以作为一种可选技术，因为此时互补金属氧化物半导体（CMOS）已经成为了主流器件。由于TTL电路在早期的电路设计中占据很重要的基础性地位，因此TTL电路广泛应用在各种小规模和中规模的功能模块设计中，而且非常耐用，并具有相对较高的运行速度和较出色的电气特性。因此，在一段时间内，TTL电路将继续作为一项重要的技术不断出现。不过，TI（Texas Instrument）公司的54/7400TTL系列成为了事实上的标准，其器件序列号和引脚经常被其他TTL制造商借鉴使用。

在实际应用中，TTL器件包含许多系列，这些器件具有不同的电路和半导体工艺参数，因此表现出不同的速度-功率特征。每个器件都会被指定一个设计组合序号和字母，用来准确确定器件和TTL的系列。以54作为起始系列代号的器件，其性能是基于军用温度范围的，该范围为-55～125℃；而74系列器件的指定温度范围为0～70℃。接着就是用来区分系列的字母，在字母后面则是用来指定器件功能的数字。例如，7486是指标准TTL系列中包含4个双输入的专用OR

门电路，而74AL86则是指具有相同功能的低功率肖特基系列电路。在这些代号中，也可能会包含用来说明器件类型的附加码和电路的版本信息。通常来说，不同TTL系列的器件是可以混在一起使用的，但必须对输出（Fanout）性能（接下来即将讨论）和噪声性能多加注意。

标准TTL电路的基本功能模块是NAND门电路，如图7-2所示。如果将输入端A或B连接到一个逻辑低电压，那么并联发射极的晶体管$Q1$将处于饱和状态，同时晶体管$Q2$的基极将处于低压情况，从而导致$Q2$处于截止状态。这样，晶体管$Q4$就缺少基极电流，从而也处于截止状态；而$Q3$通过电阻$R2$接受了基极驱动电流，从而处于导通状态，并将输出电压上拉至V_{CC}。标准TTL电路的典型输出高电压值为3.4V。另一方面，如果输入端A和B都连接高电压，那么晶体管$Q1$将转换到反相有源放大区，其集电极的输出电流将成为$Q2$的基极输入电流。这样，$Q2$就导通了，并使$Q4$的基极电压升高，直至$Q4$处于饱和状态。同时，$Q2$集电极电流在电阻$R2$上产生的电压将会降低$Q3$的基极电压，使其值低于用来维持$D1$和$Q3$的基极-发射极结的前向偏置电压。这样，$Q3$就处于截止状态，其输出端的电压就基本接近于接地电压。接下来，在正常的工作过程中，$Q2$和$Q4$将在截止和饱和状态之间转换。饱和晶体管在它们的基极区将产生一个少数载流子的聚集过程，而晶体管离开饱和区域所需的时间取决于多余的载流子能以多快的速度从基极区离开。标准的TTL工艺会将金元素作为掺杂物来产生诱捕点，这个过程必须基于少数电子的加速湮没。最新的TTL系列融合了电路的精巧设计和工艺改进设计，用来降低多余载流子的浓度并加速基极聚集电子的转移。

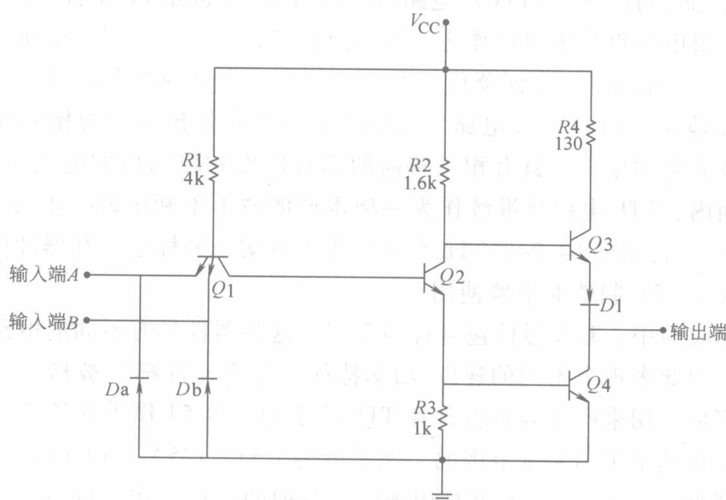


图7-2　包含两个输入端的标准TTL NAND门电路

由于内部转换机制不同，而且上拉和下拉晶体管具有不同的电流驱动能力，TTL 器件将输出电压从低电压提升至高电压时所用的传输时间将会很长（见图 7-3）。虽然在速度比较的过程中经常使用平均传输延迟（$(t_{phl}+t_{plh})/2$），但是保守的设计师在计算电路性能时会使用相对较低的传输延迟。例如，标准 TTL 中 NAND 门电路的延迟典型值为 10ns，其最大 t_{phl}为 15ns，最大 t_{plh}为 22ns。

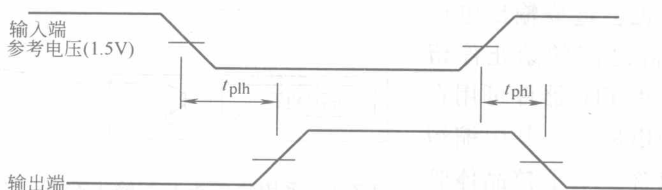


图 7-3　传输延时

当 TTL 门电路的输出较低时，其输出电压也是下拉晶体管 Q4 的集电极-发射极电压。如果从正在被驱动的电路中过来的集电极电流增加，输出电压也会升高。因此，输出电流必须被限制，以确保输出电压处于指定的逻辑低压范围之内（低于 V_{OL}）。这样输入端的数量必须定义一个最大值，这些输入端由给定输出电压低压驱动。换句话说，任何 TTL 的输出都有一个最大的低压输出端数量（fanout）。

类似地，在高逻辑电压下也有一个最大输出端数量。由于电路必须在任何逻辑电压下都能正常工作，因此这两个输出之间的较小者就可以确定整体的最大输出端数量，如下所示：

$$\text{fanout} = \text{minimum}(-I_{OH,max}/I_{IH,max},\ -I_{OL,max}/I_{IL,max})$$

式中，$I_{OH,max}$为 $V_{OH,min}$指定的最大输出电流；$I_{OL,max}$为 $V_{OL,max}$指定的最大输出电流；$I_{IH,max}$为高电压下指定的最大输入电流；$I_{IL,max}$为低电压下指定的最大输入电流。

输出计算公式中之所以出现负号，是因为输入电流和输出电流的参考方向由器件定义。大多数标准 TTL 系列中的器件在驱动其他标准 TTL 电路时，其输出端数量为 10。具有提高输出电流驱动能力的缓冲器电路可以用于很多场合，如时钟信号电路，该时钟信号必须用来驱动数量非常多的输入端。TTL 系列中，除了标准 TTL 电路外都具有各不相同的输入和输出电流要求，因此，只要是不同系列的器件混用在一起，就需要重新计算输出端数量。

常规“图腾柱”TTL 输出电路中总有一个晶体管始终处于导通状态，并将输出端的电压上拉至一个高逻辑电压，或者下拉至一个低逻辑电压。因此，两个或两个以上的输出端不能连接到相同的节点；如果一个输出端正试图驱动一个低

逻辑电压，而其他输出端正试图驱动一个高逻辑电压，那么在 V_{CC}和地之间就会产生一条低阻抗的路径。多余的电流就可能会对输出晶体管产生损害，即使不会产生损害，也会造成输出电压处于 V_{OH} 和 V_{OL} 之间的禁止停留过渡区域。有些 TTL 器件可用在集电极开路的电路中，其中缩短的输出端只包含一个下拉晶体管（见图 7-4）。两个或两个以上的集电极开路输出端可以通过一个上拉电阻连接到 V_{CC}，从而构成一个连线-AND 模拟逻辑门电路。如果所有晶体管都导通，那么共极点的电压就会被拉下来；如果所有晶体管都处于截止状态，那么共极点的电压就很高了。上拉电阻的阻值必须足够大，以确保当单个开关驱动输出低压时，I_{OL}不会超出范围；而上拉电阻的阻值又必须足够小以确保当所有驱动门电路都处于截止时，输入端泄漏电流不会将输出电压下拉至低于 V_{OH} 的程度。由于增加上拉电阻的阻值会增加连线-AND 门电路的电阻-电容（RC）时间常量，同时也会增大有效传输延迟，因此上拉电阻通常选择接近于最小值的值。在实际设计中，公式为：

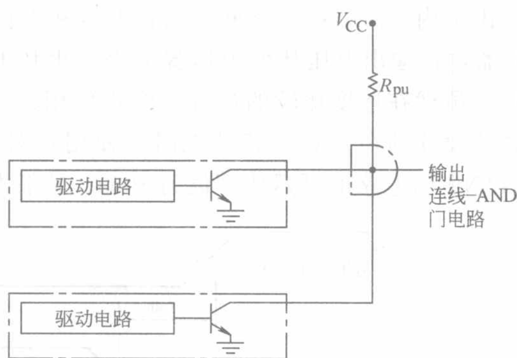


图 7-4 采用集电极开路输出端和一个上拉电阻的连线-AND 门电路

$$(V_{CC}-V_{OL,max})/(I_{OL,max}+nI_{IL,max})<R_{pu}<(V_{CC}-V_{OH,min})/(mI_{OH,max}+nI_{IH,max})$$

式中，m 为连接到 R_{pu}的开路集电极开关数量；n 为被驱动的输入端数量。

所有电流的参考方向都由器件的引脚决定。

有些 TTL 器件具有三态输出——图腾柱输出端，该输出端具有一个特殊的功能，即两个输出晶体管可以在一个输出使能信号的控制下截止；当输出使能触发时，器件的输出功能就会有效断开输出引脚与内部电路的连接；当使能没有触发时，两个晶体管都会截止。许多三态输出端可以连接到一个共用的总线，而且同时只有一个输出端被使能，从而在总线上选择性地设置逻辑电压。

纵观 TTL 的发展历程，人们提出了很多方案来实现更快的速度、更低的功耗。例如，54/74L 系列的器件通过使用更高阻值的电阻来合并重新设计的内部电路，以此降低电源电流，并将功耗减小至标准 TTL 电路的 1/10。但同时速度也被减慢，其延迟时间变为原来的 3 倍。相反，54/74H 系列可以提供比标准 TTL 电路更高的速度，但代价是增加了功耗。以上这两种系列的电路都已经被弃用了，取代它们的是各种样式的肖特基 TTL。

如前所述，TTL晶体管可以被驱动至饱和状态。与使用金元素掺杂物来加速少数载流子的衰减不同，肖特基晶体管限制了集电极-基极结的前向偏置电压，因此，肖特基晶体管通过并联集电极-基极结与肖特基（金-半导体）二极管来加速基极少数载流子的聚集，如图7-5所示。当NPN晶体管开始进入饱和状态时，肖特基二极管就处于前向偏置状态。这样，集电极-基极结的前向偏置电压被控制在了0.3V，使得晶体管处于固定饱和状态之外。肖特基TTL（S系列）的运行速度比标准TTL电路快3倍，而同时其功耗则是标准TTL电路的1.5倍。

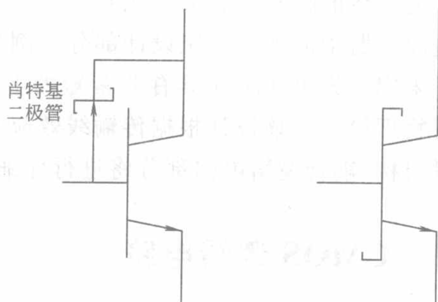


图7-5　肖特基晶体管结构及其电路符号

低功耗肖特基TTL（LS系列）将一个改进的输入电路设计与肖特基晶体管合并起来，以此来将功耗降低至标准TTL电路的1/5，同时维持与原来相同或比原来更高的运行速度。LS逻辑电路中重新设计的输入电路改变了输入逻辑开启电压，并导致噪声容限略微减小。

另外一个TTL系列就是“Fairchild先进肖特基TTL（Fairchild Advanced Schottky TTL，FAST）”。Fairchild融合了电路在工艺技术上的各种改进，这种改进是指降低了结电容并提高了晶体管的运行速度。这使得该系列器件的运行速度变成了标准TTL系列元件的5倍，同时具有更低的功耗。

顾名思义，先进肖特基（Advanced Schottky，AS）和先进低功耗肖特基（Advanced Low-power Schottky，ALS）系列使用了改进的制造工艺来提高晶体管的转换速度；这些系列器件的运行速度是S和LS系列元件的2～3倍，而功耗只有S和LS系列器件的1/2。尽管标准TTL、肖特基（S）和低功耗肖特基（LS）系列器件仍然在批量生产，但是更新的先进肖特基系列（AS、ALS、F）可以提供更好的速度-功率乘积，因此更受新型电路设计的欢迎。

在比较不同逻辑系列的速度时，必须记住一点，就是制造商提供的说明规范对于不同的系列对应了不同的负载电路。例如，标准TTL的传输延迟是在一个负载电路中测得的，在该负载电路中有一个15pF电容和一个并联的400Ω电阻；而FAST的传输延迟是在包含50pF电容和500Ω电阻的负载电路中测得的。由于有效延迟随着负载电路电容的增加而成线性增加，因此，直接比较原始制造商的数据是毫无意义的。制造商手册中的校正曲线说明了传输延迟随负载电容变化的关系。

当输出电压从一个逻辑状态变化到另一个逻辑状态时，电源电流会出现瞬变

现象。在 V_{CC}和 TTL 逻辑电路的接地线之间放置了去耦电容，以提供过滤功能并使沿电源线的瞬间电流最小化并减少噪声的产生。去耦电容的分布原则可以参考制造商手册中的数字系统设计部分，例如 Barnes（1987）。

采用了先进 TTL（具有非常短的上升时间，如 FAST 和 AS）的电路板需要进行物理设计，该设计根据传输线效应来减小波形误差。关于这些内容，在稍后的发射极-耦合逻辑电路部分将进行详细讨论。

7.3 CMOS 逻辑电路

1962 年，Fairchild 半导体公司的 FrankWanlass 注意到了增强型 NMOS 和 PMOS 晶体管可以堆放成图腾柱形式，以形成一个非常简单的反相器（见图 7-6）。如果输入电压接近于 V_{DD}，那么 NMOS 下拉晶体管的沟道将会被扩展，而 PMOS 晶体管将处于截止状态。低沟道阻抗 NMOS 和高沟道阻抗 PMOS 的分压器效应产生了一个低输出电压；换句话说，如果输入电压接近于接地电压，那么 NMOS 晶体管将处于截止状态，而 PMOS 晶体管的沟道将会被扩展，从而将输出电压上拉至高值。通过添加并联下拉晶体管和串联上拉晶体管，就可以很容易得到逻辑门电路。当一个 CMOS 门电路处于静态时，惟一存在的电流是通过截止晶体管的微小漏电流加上驱动输入端所需的电流。由于连接到 CMOS 门电路的输入端基本上都是电容，这样，输入端电流就非常小（属于 100pA 级别）。因此，CMOS 门电路的静态电源电流和功耗就远比 TTL 门电路的小。另外，由于早期的离散 CMOS 逻辑电路的运行速度远比 TTL 慢，因此，CMOS 经常应用在功率守恒比性能更重要的领域中。

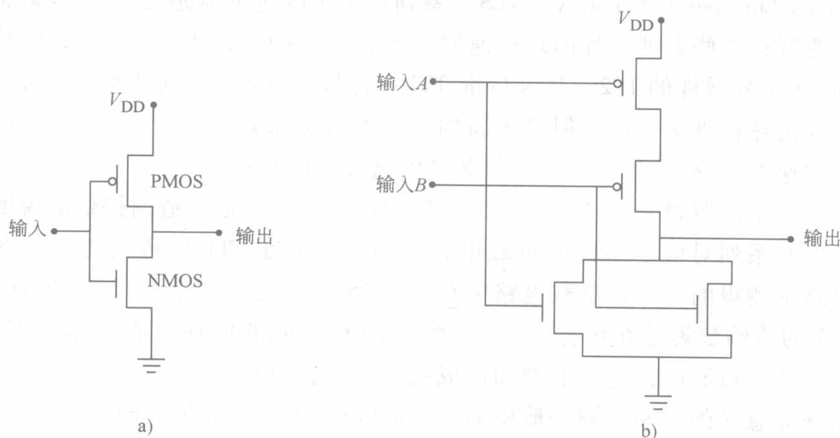


图 7-6　基本 CMOS 逻辑电路

a）互补 MOS 反相器　b）双输入端 CMOS NOR 门电路

在 TTL 逻辑电路中，高输出电压和低输出电压以及输入转换开启电压（一个特殊开关输入端将其作为非零的电压值）主要由 PN 结的前向偏置电压确定。由于这些电压并不随电源电压成比例变化，因此，TTL 电路工作在一个严格限制的电压范围之内（5.0±10%）V。但是，在全 CMOS 系统中，输出电压由分压器效应决定。由于非动态晶体管沟道的截断阻抗远比动态晶体管沟道的大，因此，CMOS 的逻辑高电压和低电压都接近于电源电压值（V_{DD}和接地电压），而且随电源电压成比例变化。在设计时，输出晶体管通常在导通状态下具有匹配的沟道阻抗。因此，当电源电压处于晶体管开启电压（对于现代 CMOS 晶体管，约为 0.7V）之上时，输入转换电压也会随电源电压成比例变化，约为 V_{DD} 的一半。与 TTL 不同，CMOS 器件可以工作在很广的电压范围（3～15V），尽管低电压会减小抗扰性和运行速度。在 20 世纪 90 年代中期，集成电路制造商生产出了新一代的 CMOS 器件，并对该器件在 3V 范围内的工作性能进行了优化处理，主要用于由两节干电池供电的便携式产品中。

只需要 4 个晶体管的基本 CMOS 门电路在芯片上所占的面积远小于包含 11 个二极管和双极性晶体管以及 6 个扩散电阻的 LS TTL 门电路在芯片上所占的面积。由于这个原因，再加上更高的静态功耗，使 TTL 电路失去了成为超大规模集成电路设计可行性技术的资格。在 20 世纪 80 年代中期，CMOS 电路快速舍弃了更早期的 NMOS 集成电路，并迅速成为复杂数字集成电路设计中的主流可选技术。

稍晚一些出现的离散 CMOS 集成电路（RCA's 4000 系列是标准，许多制造商都生产相应的器件）使用铝作为栅极材料。铝金属栅极在漏极和源极区域形成之后才累积形成。为了确保栅极即使在出现最大隔离层排列误差情况下也能覆盖整个沟道，在栅极和源极与漏极之间必须存在重叠的部分。但是，过大的寄生电容会导致晶体管（栅极）速度变慢。在 20 世纪 80 年代中期，许多制造商都生产先进的离散 CMOS 逻辑电路，该逻辑电路基于自动排列的多晶硅栅极工艺，该工艺通常用于超大规模集成电路（IC）中。离散硅栅元件通常会和 TTL 器件在功能上产生竞争，而且具有类似的序号设计（例如，74HCxxx）。由于具有类似的静态功耗优势（1mW 和 0.6mW/每个栅极），硅栅 CMOS 的运行速度远比金属栅极的快（8ns 和 125ns 的典型栅极传输延迟），因此，硅栅 CMOS 成了新型离散逻辑电路设计中的竞争技术之一。现代 CMOS 器件中可以分为两种截然不同的类型：HC 高速 CMOS 和 AC 或 C 先进 CMOS。

尽管 74HC 和类似系列的器件可能会效仿 TTL 器件的功能，而且具有匹配的引脚排列，但是它们不能和 TTL 器件混在一起使用。如前所述，CMOS 的输出电压相比 TTL 电路的输出电压更接近于电源电压，其设计通常会将这个特征考虑在内；而且当栅极由典型的 TTL 高逻辑电压驱动时，不会产生固态输出电压。

在生产专用的先进 CMOS 器件时会对输入端进行改进，尤其是和 TTL 电路（如 HCT 系列）配合使用时。这些器件的功耗比正常硅栅 CMOS 稍微高一点。

CMOS 晶体管的栅极与沟道是通过一层 SiO_2 进行隔离的，该隔离层可能只有数百埃的厚度（1Å（埃）$=10^{-10}$m）。如此薄的介电层（隔离层）很容易就被区区 40～100V 的电位差破坏。如果不采取预防措施，正常的人体静电就可以损坏 CMOS 器件。尽管通过合适的处理工艺（Matisof，1986）可以将由静电放电（Electro Static Discharge，ESD）产生的损害减小到最低程度，但是 CMOS 器件还是在所有输入端上都添加了 ESD 保护电路（见图 7-7）。图 7-7 中的输入二极管用来将栅极电压限制在一个二极管压降的值，要么高于 V_{DD}，要么低于接地电压；而图中的电阻同时会限制输入端电流并延长快速脉冲的上升时间。

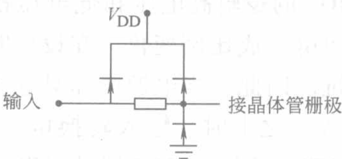


图 7-7 典型的 CMOS 输入保护电路

如图 7-8 所示，CMOS 反相器与栅极跨接在寄生双极性晶体管上，这种结构形成了一个硅控整流器（Silicon Control Rectifier，SCR）。假设连接到输出引脚的外部电路将引脚电压拉下来并低于接地电压，当引脚电压接近 −0.5V 时，寄生 NPN 晶体管的基极-发射极结就会开始产生前向偏置，从而晶体管就会导通；继而产生的集电极电流主要由外部电路将输出端电压拉下来的过程决定。该集电极电流将会通过 R_{well} 并降低寄生 PNP 晶体管的基极电压。一个足够大的电流对 PNP 晶体管的基极-发射结进行前向偏置时，将会导致 PNP 晶体管的集电极产生

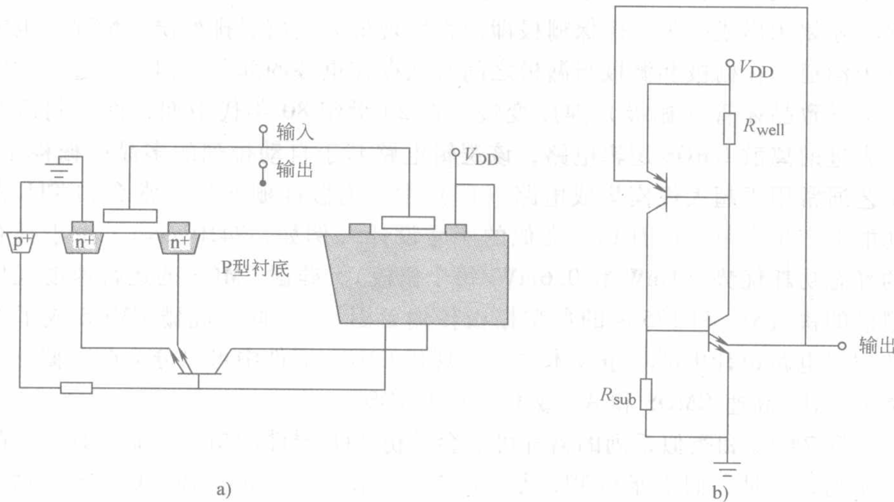


图 7-8 CMOS 寄生 SCR
a）寄生双极性晶体管的横截面 b）等效电路

电流并通过 R_{sub}，从而有助于维持 PNP 晶体管的前向偏置电压。接着，SCR 就进入了一个再生状态，并快速呈现出一个稳定状态，该稳定状态下，即使初始驱动电源被移走，两个晶体管仍然可以保持导通。在这种“锁定”（Latchup）情况下，电流将直接从 V_{DD}流至接地点；这样，CMOS 门电路的正常运行就被打乱了，而且有可能被永久性破坏。要产生“锁定”现象，NPN 晶体管和 PNP 晶体管的电流增益的乘积必须比其中之一大，输出引脚必须由低于 -0.5V 或高于 V_{DD} +0.5V 的电压来驱动，该电压必须具有足够大的电流来触发再生过程，而且 V_{DD}必须提供一个持续的电流来维持 SCR 处于导通状态。尽管“锁定”隐患不可避免，但是 CMOS 制造商会将输入端和输出端设计成可以抵抗“锁定”的电路，在这个设计过程中，制造商会使用可以减小寄生晶体管增益和降低衬底及 P 阱电阻的配置。不过，CMOS 用户必须确保输入端和输出端在驱动时不会超出正常的工作范围。另外，将电源电流限制在有效电流水平之下也是避免“锁定”损害发生的一种方法。

驱动 TTL 输入端的 CMOS 输出端的数量会受到静态电流驱动要求的限制，而且在晶体管-晶体管逻辑电路中必须考虑在内。在全 CMOS 系统中，静态输入电流就是指通过输入 ESD 保护二极管的泄漏电流，其大小属于 100pA 级别。因此，驱动 CMOS 的 CMOS 静态输出端数量是如此之大以至于变得没有任何意义了。相反，输出端数量会受到随驱动栅极数量增加而退化的速度的限制（Texas Instruments，1984）。如果假设所有被驱动的输入端都是相同的，那么一个全 CMOS 系统就可以由一个等效电路来描述，如图 7-9 所示。并联输入电阻 R_{in} 属于 60MΩ 的级别，在稍后的计算过程中可以忽略。上拉电阻 R_{pu} 的阻值可以由制造商的数据（$V_{DD}-V_{OH}$）/（I_{OH}）（~50Ω）计算得到。输出电压将会随时间常数 $nR_{pu}C_{in}$从低至高呈指数变化。其中，n 是指实际输出端数量。由于低输出电压接近于零，因此，达到高逻辑电压所需的时间为

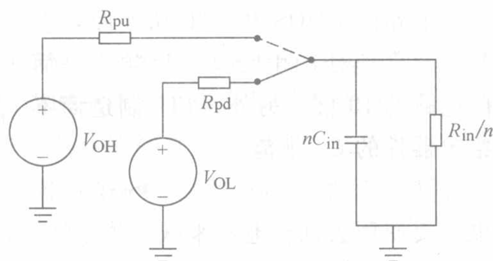


图 7-9 CMOS 输出端的等效电路（用来驱动 nCMOS 输入端）

$$t=nR_{pu}C_{in}\ln(1-V_{IH,min}/V_{OH})$$

这样，一旦指定了一个全 CMOS 系统的最大可行上升时间（也就是最大有效传输延迟），就可以计算动态输出端数量 n 了。尽管 CMOS 逻辑电路的静态功耗非常小，但是那些正在改变逻辑状态的输出端在每次对加载在输出端上的电容进行充放电时将会消耗电源功率。另外，输出端的上拉和下拉晶体管在每个转换

周期内都会在一定程度上同时导通一段很短的时间，并产生一个过渡供电电流，该电流的值超出了电容充电的需求范围。因此，一个 CMOS 开关的总体功耗如下所示：

$$P_{\text{total}} = P_{\text{static}} + C_{\text{load}} V_{\text{DD}}^2 f$$

式中，P_{total}为整体功耗；P_{static}为静态功耗；C_{load}为整体等效负载电容，$C_{\text{load}} = C_{\text{pd}} + C_{\text{ext}}$；$C_{\text{pd}}$为功耗电容，它是一个制造商规范，代表了等效内部电容（考虑过渡转换电流效应）；C_{ext}为整体并联等效负载电容；V_{DD}为供电电源电压；f为输出端工作频率。

HCT 和 ACT 器件包含了一个额外的静态功耗，该功耗与 TTL 开始被驱动的输入端数量成正比（$n\Delta I_{\text{DD}} V_{\text{DD}}$，其中 ΔI_{DD}由制造商指定)。CMOS 的平均功耗与工作频率成正比，而 TTL 电路的功耗几乎和工作频率（低于 1MHz）无关；当超过 1MHz 时，动态功耗会变得非常大（见图 7-10)。直接比较数字系统中 CMOS 和 TTL 之间的功耗是很困难的，因为不是所有输出端都随相同的功率变化。另外，CMOS 电路的相对优势随元器件的集成度增加而增加；中等复杂的译码器电路的 CMOS 和 TTL 功率曲线会在某个频率值处产生交叉，该频率是简单门电路的 10 倍。另外，TTL 制造商不会给出器件的 C_{pd}规范。

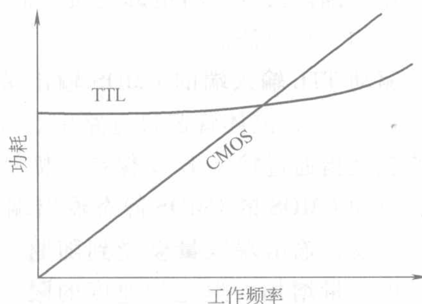


图 7-10 TTL 和 CMOS 逻辑器件的功耗-工作频率关系图

如同 TTL 器件一样，CMOS 逻辑器件也需要使用去耦合电容来减小供电线上过渡电流尖峰信号产生的噪声。尽管这些去耦电容可以减小并联信号线路上的噪声静电耦合，但是它们不能消除一个相加性干扰，该干扰在先进 CMOS 电路和先进肖特基电路中都可能产生，主要原因是供电线上过渡电流尖峰信号产生的电压。多路反相器 IC 如图 7-11 所示，其中 5 个输出端正同时从低电压转换到高电压，剩下的 1 个输出端仍保持低电压。根据法拉第定律（Faraday's law)，转换电流尖峰信号将会在 L 上产生一个过渡电压，L 代表了内部封装线和外部接地线的寄生感应系数。而过渡电压也会将不变的输出端电压提升到高于接地参考电压的水平。一个足够大的电压尖峰信号可能会使信号 A 超过输入端开启电压，并会导致后续逻辑器件的操作出现错误。电源线和接地线的精细设计可以减小外部感应系数，同时还可以限制电源电压下垂和接地端跳动。其他的有效技术还包括独立缓冲器中同步时钟的使用以及将高驱动信号（例如时钟线）分割成独立的缓冲包，这些独立的缓冲包与复位信号和其他异步缓冲器信号相互隔离。

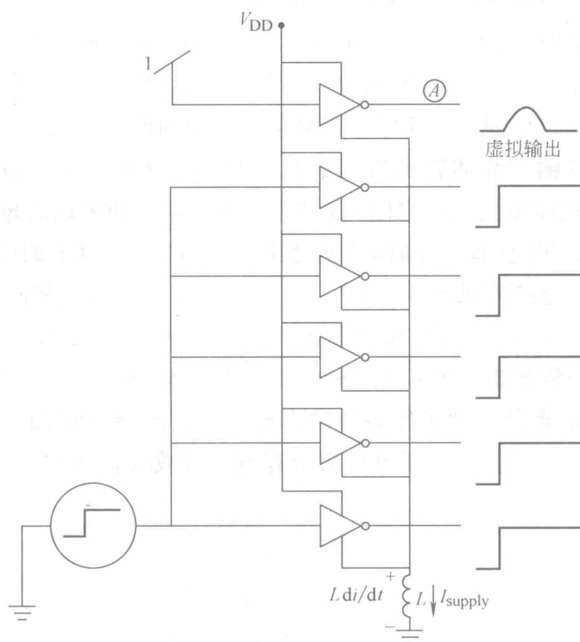


图 7-11　由接地线感应系数引起的输出端跳动

前文提到过很多次，逻辑电路的有效传输延迟与其输出端的电容性负载成正比关系。由小规模和中规模逻辑电路构成的系统具有很多集成电路输出端，这些输出端具有负载电容，该负载电容是由印制电路线互连引起的。这就是使用大规模集成电路的主要推动力：片上互连部分的电容小得多，从而具有更高的运行速度。因此，相对于由较慢的技术制造的高集成电路系统，由快速小规模逻辑器件构成的相同系统其运行速度更快。

双极性 CMOS（Bipolar CMOS，BiCMOS）逻辑是超大规模集成电路中的一种混合技术，该技术在同一块芯片上合并了双极性和 MOS 晶体管，用来实现大部分的电路集成度并节省 CMOS 的静态功率，同时还可以提供很强的电气特性和更高级的双极性逻辑器件电流驱动能力。BiCMOS 还可以利用双极性晶体管制造混合系统的线性部分，或利用 CMOS 制造混合系统的逻辑部分。在同一块芯片上制造高质量的双极性晶体管（如 MOS 器件）是一件很困难的任务，也就是说 BiCMOS 电路的成本更高，同时性能也不是最理想。

7.4　发射极耦合逻辑电路

发射极耦合逻辑（ECL）电路是运行速度最快的离散逻辑电路，同时也是最

古老的电路之一，可以追溯到1962年Motorola的MECLI电路。改进型的MECLⅢ电路最早出现在20世纪60年代后期。如今，10和100K ECL系列仍然在超高速电路中使用。不同于TTL和CMOS电路，发射极耦合逻辑电路使用了差分放大器配置作为基本拓扑结构（见图7-12）。如果所有连接到栅极输入端的电压都低于参考电压，那么相应的输入晶体管都将处于截止状态，从而整个偏置电流就会都通过Q1，其输出端的电压就会通过电阻Ra被上拉至V_{CC}。如果所有输入端的电压都高于参考电压，那么相应的输入晶体管就会全部导通，并将Q1的电流切换到输入晶体管，同时将输出电压降低至$V_{CC}-I_{bias}R_a$。如果集电极电阻和偏置电流选择恰当，那么晶体管就永远不会进入饱和状态，而且切换速度也会非常快。例如，MECL 10K门电路的典型传输延迟只有0.2ns。为了减小输出阻抗并提高电流驱动能力，实际应用中的ECL器件包含了发射极输出放大器输出端（见图7-12b），同时还包含了完整的参考电压值，该参考电压值带有与差分放大器相匹配的温度系数。

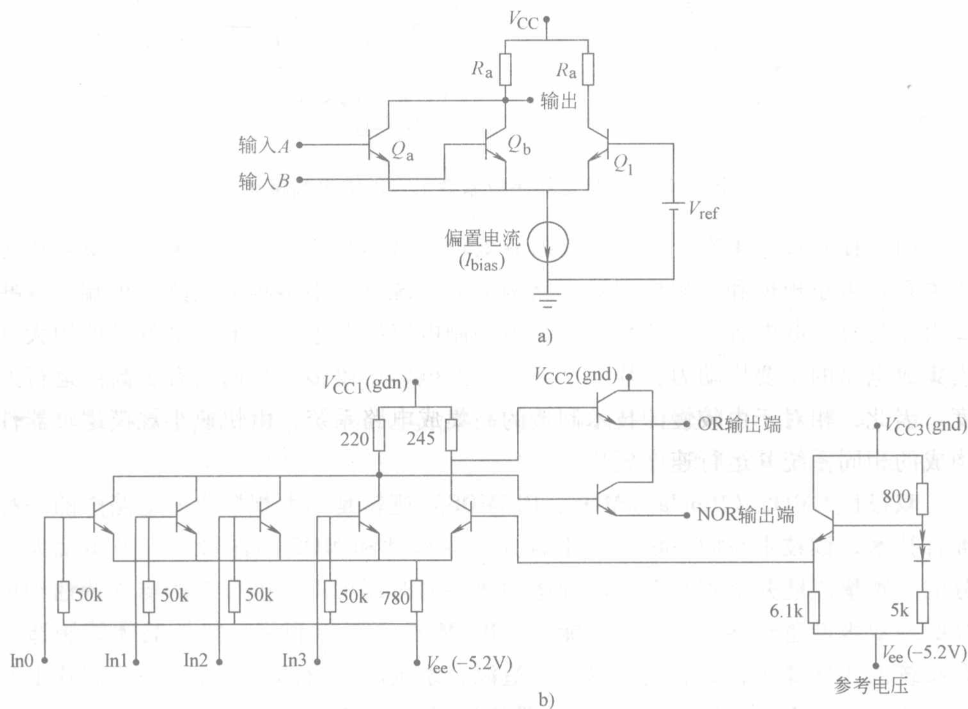


图7-12 发射极耦合逻辑（ECL）门电路

a）ECL NOR门电路简化示意图 b）经济型ECL门电路

我们可以注意到，ECL电路正常工作时其集电极接地，发射极电流源连接到负电源电压。从电源电压波动的角度来看，电路正常工作时就如同一个基本放大器，并在输出端处减弱了电源电压的波动。TTL和CMOS电路中的电源电流在器件

转换时会产生尖峰信号；在 ECL 门电路中，偏置电流保持不变，只是从一个晶体管转移到另一个晶体管。越低的逻辑电压摆动，就越容易减小噪声的产生。尽管 ECL 逻辑电路的噪声容限比 TTL 和 CMOS 都要低，但是 ECL 电路产生的噪声更少。

发射极耦合逻辑电路具有很短的传输延迟和很快的上升时间。如果互连路径上产生的延迟比上升时间多一半，那么互连部分的传输线特性就必须考虑在内。非终端线的反射会对波形造成影响，使其扭曲，并伴随激振。等待这种反射衰减到可接受值的过程，会导致传输迟延增加。最大的开路线长度取决于驱动输出端的数量、传输线特征系数以及传输线的传输速度。Motorola 公司的数据手册指出，MECLⅢ器件可以容忍的非终端线长度只有 0.1in㊀，该器件的实际输出端数量为 8，并连接了小口距的印制电路线（100Ω 的微带）。在高速 ECL 设计中，物理互连部分必须使用合适的终端传输线来制造（见图 7-13）。关于数字系统中传输线的扩展方法，可以参考制造商的说明（Blood，1988）或（Rosenstark，1994）。在实际全 ECLⅢ设计中必须使用传输线，这个要求导致了 1971 年 MECL 10K 系列器件的诞生。该系列器件具有 2ns 的传输延迟和 3.5ns 的上升时间，这种被延缓的上升时间说明可以使用更长的非终端传输线（在前文提到的配置中，Motorola 采用 2.6in）。这样，许多设计都可以利用 10K ECL 来实现，而无需传输线互连。这种使用上的方便和器件种类上的丰富性使得 10K ECL 及其派生类型成为了 ECL 系列的主流器件。但是，我们应该注意到，尽管传输线互连传统上与 ECL 相关，但是先进 CMOS 和先进肖特基 TTL 电路具有很小的上升时间，而

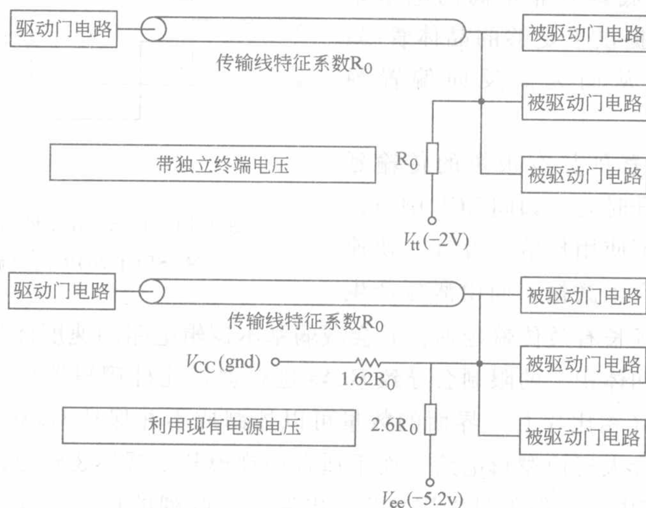


图 7-13　并联终端 ECL 互连电路

㊀ 1 英寸（in）=0.0254m。——译者注

且经常应用于超高速系统中。在任何高性能系统的物理设计中，无论使用什么逻辑系列都必须考虑传输线效应。

相比 TTL 和 CMOS 技术来说，尽管 ECL 经常作为一种“功耗匮乏”技术，但是在超高工作频率下时就不一定是这样了。在高频条件下，TTL 和 CMOS 的功耗基本上随频率线性增加，而 ECL 的功耗保持不变。当频率达到 250MHz 时，CMOS 的功耗就超过了 ECL 的功耗。

我们知道，在晶体管出现之前，砷化镓（GaAs）构成了半导体化合物，该化合物的电子迁移率比硅晶体高很多。但是，空穴的迁移率比硅晶体慢。因此，利用电子作为电荷传输载体的单极 GaAs 晶体管的运行速度就比硅晶体高，而且可以用来生产具有超短传输延迟的逻辑电路。由于 GaAs 没有天然的氧化层，因此，很难用来制造高质量的 MOS 场效应晶体管（MOSFET）。首选的 GaAs 逻辑晶体管是金属-半导体场效应晶体管（MESFET），该场效应晶体管实际上是一个结型场效应晶体管，其栅极由金属-半导体（肖特基）结形成，而不是 PN 二极管。图 7-14 给出了一个 NOR 门电路的示例，在该电路中两个增强型 MESFET 管作为下拉晶体管，一个耗尽型 MESFET 作为上拉动态负载。在室温下，GaAs 的衬底具有非常高的电阻率（即半绝缘），提供了足够的晶体管-晶体管绝缘性，从而无需反向偏置绝缘结。

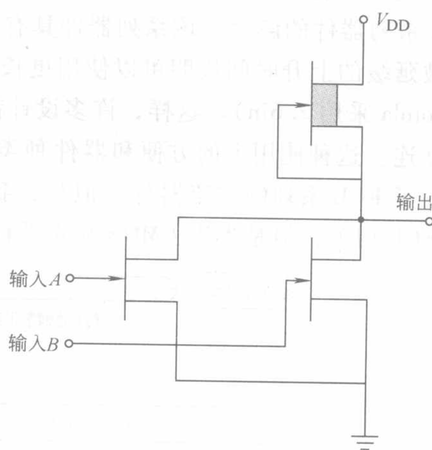


图 7-14　GaAs 增强型/耗尽型 MESFET NOR 门电路

GaAs 逻辑器件具有很短的传输延迟和很快的上升时间。如同 ECL 中讨论的一样，如果不使用传输线互连，快速逻辑信号就会产生激振，而由激振产生的波形扭曲会延长有效传输延迟，并会减弱基本逻辑电路的速度优势。此外，复杂的制造工艺和体积上的限制会导致 GaAs 逻辑器件比硅逻辑器件贵 10～20 倍。通过限制片下（相比片上）界面的数量可以控制成本并保持 GaAs 的速度优势，这种技术更适合大规模逻辑电路，而不适合中规模和小规模逻辑电路。销售商提供的 GaAs 数字电路一般都是以用户集成电路或门阵列的形式出现的，并由用户根据所需的逻辑电路来指定门阵列的最终电路排版图形样式。综上所述，销售商可以提供专用的 GaAs 集成电路，这些电路中包含了 20000 个门，而且最差的内部开关延迟也只属于 0.07ns 的级别。

7.5 可编程逻辑电路

尽管传统的离散逻辑器件是实现数字系统的重要方式之一，但是目前各种各样的可编程逻辑器件也已经越来越受欢迎了。从广义上来说，可编程逻辑器件是指那些包含常见逻辑模块阵列的大规模和超大规模集成电路，在逻辑模块中包含了用户可配置互连部分。在对器件进行编程时，由用户来定义每个逻辑模块和逻辑模块之间互连部分的功能。根据使用的可编程逻辑器件的集成度，单个可编程逻辑包可以用来代替成批的传统小规模和中规模集成电路。由于大部分的互连部分都在可编程逻辑器件芯片上，因此，可编程逻辑器件设计可以提供比传统逻辑电路更好的速度和可靠性。可编程逻辑器件的设计可以采用电子设计自动化（Electronic Design Automation，EDA）工具进行，EDA 工具可以根据功能来描述自动生成器件的编程信息。功能性描述可以按照如下形式进行：逻辑图、逻辑方程式或软件描述语言。EDA 工具还可以执行以下任务，如逻辑合成、电路简化、器件选择、功能性仿真以及计时/速度分析。电子设计自动化和器件的可编程使得数字系统的设计与物理实现相比传统逻辑电路变得更快了。

尽管有时区分得不很精确，但是可编程逻辑器件仍然可以分为两个主要类型：可编程逻辑器件（PLD）和现场可编程门阵列（Field Programmable Gate Arrays，FPGA）。PLD 又可分为包含很多内部逻辑模块的复杂可编程逻辑器件（Complex Programmable Logic Device，CPLD）和简单 PLD。虽然小型 PLD 的通用结构是十分标准的，但是 FPGA 和 CPLD 仍在不断发展，各个制造商的产品结构也大不相同。实际上，可编程逻辑器件的概念覆盖了很广的范围，包含了从几个门阵列构成的简单可编程阵列逻辑（PAL）到可以提供数千个可用开关等效电路的 FPGA。

1. 可编程阵列逻辑

可编程阵列逻辑（PAL）最早由 Monolithic Memories 公司推出。在最简单的结构中，PAL 由一长列的逻辑 AND 门电路组成，该逻辑 AND 门电路由输入信号的源码（非反向）和补码（反向）来进行反馈。任何一个逻辑 AND 门电路都可以由任意一个输入信号反馈。逻辑 AND 门电路的输出端通过固定线路与 OR 门电路连接，OR 门电路用来驱动 PAL 的输出端（见图 7-15）。通过对 AND 输入端进行编程，PAL 用户可以实现任何逻辑功能，该逻辑功能提供的乘积项不会超过 OR 门电路的输入端数量（fanin）。现代 PAL 合并了许多宏单元，这些宏单元中包含一个可编程反相器、一个 AND/OR 组合逻辑电路以及一个可编程触发器，该触发器在寄存器输出端需要时可以使用。附加的 AND 门电路通常用于时钟电路和复位触发器电路。宏单元通常包含一个可编程三态缓冲器，该缓冲器用来驱

动时钟输出和可编程门电路，可编程门电路可以将宏单元输出反馈到AND-OR阵列，用来开发组合逻辑电路（产生很多乘积项）或用来实现状态机。

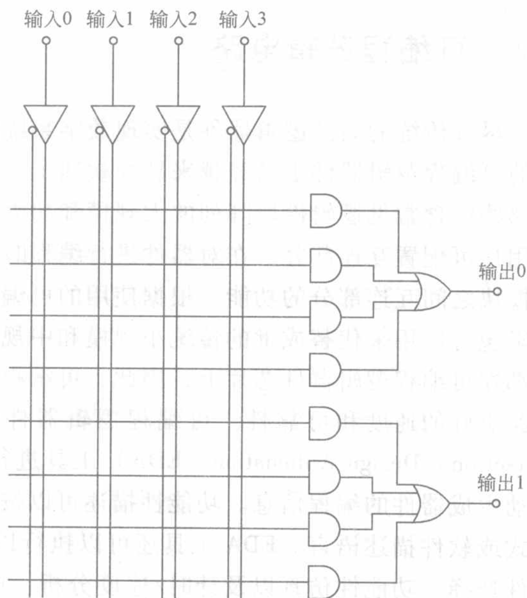


图 7-15 简化的小型 PAL 器件

PLD是使用双极性或CMOS技术制造的。尽管CMOS器件的运行速度已经非常快了，但是双极性PAL可以提供更高的运行速度（输入端至输出端的组合延迟 <5ns），但同时也具有更大的功耗。不同的制造商对互连和逻辑选项的编程方式是不一样的。在某些情况中，电路必须通过熔固可熔片或抗熔器（抗熔器是一种开路电路，当被熔断时就变成了一个低阻抗电路）来进行物理改进。CMOS可编程器件通常使用基于可擦除可编程只读存储器（Erasable Programmable Read Only Memory，EPROM）的互连单元。基于EPROM的器件可进行编程、擦除以及再编程。尽管再编程功能在器件的开发阶段很有用，但是在生产中却不是很重要，而且基于EPROM的PLD售价更便宜，而且是一次性的可编程版本。尽管没有占据主流市场，但是可再编程PLD还是很有用处的。当应用在数字系统中时，PLD必须进行输出端数量计算，就像离散TTL或CMOS逻辑电路中一样。另外，快速PLD同样需要考虑传输线效应。

尽管可以手动开发简单PAL的程序表，但实际上所有的设计都是通过使用由PLD销售商或第三方提供的EDA工具来进行的。尽管计算机方法在一定程度上自动处理了设计过程，但是它仍然允许部分用户对资源分配进行控制；用户必须对指定器件的结构和程序选项有一个非常彻底的了解，以便使最后的设计达到最优化。

2. 可编程逻辑阵列

现场可编程逻辑阵列（FPLA）类似于PAL，不同之处是，FPLA中从AND媒介输出端到OR输入端的连接部分也是可编程的，而不是固定线路（见图7-15）。类似于简单的PAL，FPLA在生产时都有寄存器输出端，即在OR门电路和输出引脚之间有一个触发器。尽管比PAL更复杂，但事实上FPLA是在PAL之前生产出来的。可编程的AND/OR互连单元使得FPLA比PAL更加灵活，因

为乘积项（AND 门电路输出端）可以驱动不止一个 OR 输出端，而且任何数量的乘积项都可以用来开发特殊输出功能。但是，可编程互连单元所在的附加层消耗了芯片面积，因此，成本更高。互连单元的级别越高，编程的程序就越复杂。但是，FPLA 最大的缺陷是运行速度比较慢。可编程互连单元添加了很多串联电阻，因此降低 AND 和 OR 门电路之间信号传输的速度，使得 FPLA 的运行速度变得比 PAL 慢了。由于 PAL 的速度更快，更加易于编程，而且对大多数应用来说更加灵活，因此 PAL 比 FPLA 更受欢迎。如今，市面上大多数的 FPLA 都具有寄存器输出端（也就是说在 OR 门电路和宏单元输出端之间放置了一个触发器），该输出端作为反馈连接到 AND/OR 阵列。现场可编程逻辑序列发生器通常用来实现小型状态机。

3. FPGA 和 CPLD

在复杂 PLD（CPLD）和 FPGA 之间，没有非常明确的分界线。不同制造商的产品之间不仅结构不同，而且互连单元编程的方法也不同。CPLD 通常是指那些由大型 PAL 阵列构成的器件，该器件将宏单元输出端的固定延迟反馈到 AND-OR 阵列（见图 7-16）。CPLD 的延迟时间比 FPGA 的延迟时间更加容易预测，这使得它更易于设计。CPLD 和 FPGA 之间更重要的一个区别是逻辑单元中组合逻

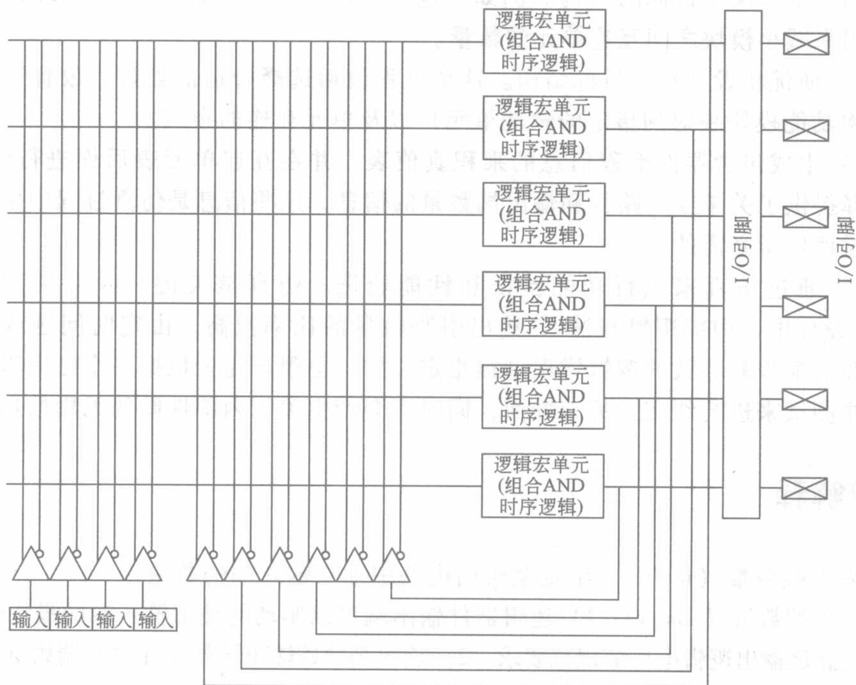


图 7-16　简化的复杂可编程逻辑器件结构

辑与连续逻辑的比例。CPLD与PAL类似的结构说明了每个逻辑宏单元中的触发器都是由多个逻辑乘积（“与”操作的结果）的总和反馈的。乘积可以包括许多输入信号和反馈信号。因此，CPLD非常适合诸如状态机的应用，状态机中包含了很多具有标量参数的逻辑功能。而另一方面，FPGA趋于向单元功能模块的方向发展，单元功能模块中包含一个触发器，该触发器由一个逻辑功能驱动，而该逻辑功能大约包含3个变量；因此，FPGA更适合需要很多寄存器的情况，这些寄存器由相对简单的功能反馈。

FPGA和CPLD在每个封装中都集成了比传统离散逻辑和小型PLD更多的逻辑器件，这使得计算机辅助设计变得更加基础，其总体设计流程如下所述：

1）通过示意图、布尔方程式、逻辑运算真值表或者高级编程语言向EDA系统描述系统的功能。每个子系统可以分离出来进行单独描述，在描述时使用最适合该子系统的描述方式，然后再将各子系统合并起来。当系统变得越来越复杂时，高级语言就会变得越来越有用。设计中后面的步骤就可以由EDA软件自动完成了。

2）将功能性描述转换到等效逻辑单元及其互连单元中去。

3）优化第2步中的结果，并得出一个适合执行的结果。优化过程及其策略的目的完全取决于目标的结构。例如，包含受限互连单元的FPGA的EDA工具可以用来减小模块之间互连单元的数量。

4）使优化设计符合目标结构。这个过程包括选择合适的器件、放置逻辑单元（将功能模块映射到指定的物理单元）以及单元互连的布线。

5）生成包含器件编程信息的乘积真值表，并在互连单元表后面进行注释。该注释提供了关于互连路径和输出端数量的信息，这些信息是仿真过程中生成精确定时评估所必需的。

6）通过仿真来执行功能验证和性能验证。仿真是关键，因为在FPGA/CPLD设计中，内部逻辑与输入-输出引脚数量的比率很高。由定时问题或编译缺陷和适配步骤导致的逻辑错误是很难定位的，也很难完全根据一个已编程器件的性能测试来进行纠正。另一方面，同时还会产生关于内部性能的无用信息。

名词解释

输入端数量（fanin）：给定逻辑门电路中独立输入端的数量。

输出端数量（fanout）：1）逻辑器件输出端可以驱动的类似输入端的最大数量，同时还满足输出逻辑电压的规范要求。2）连接到特殊输出端的实际输入端数量。

锁定（Latchup）：CMOS电路中不完美的运行状态。在该状态中，其寄生SCR在电源线之间产生了一个低阻抗路径。

噪声容限：一个输出逻辑高（低）电压和输入逻辑高（低）电压之间的差异。该输出逻辑电压由一个合适的功能输出端产生，而输入逻辑电压由被驱动的输入端需求定义。噪声容限提供了一定程度的抗信号扭曲的能力，该信号扭曲的程度低于噪声容限。

上升时间：一个逻辑信号从一个静态电压转换到另一个静态电压时所需要的时间。通常，上升时间从10% ~90%的终端信号值中测得的。

参考文献

[1] Altera. 1993. *Data Book.* Altera Corp.，San Jose，CA.

[2] Alvarez，A. 1989. *BiCMOS Technology and Applications.* Kluwer Academic，Boston，MA.

[3] Barnes，J. R. 1987. *Electronic System Design：Interference and Noise Control Techniques.* Prentice-Hall，Englewood Cliffs，NJ.

[4] Blood，W. R. 1988. *MECL system Design Handbook.* Motorola Semiconductor Products，Phoenix，AZ.

[5] Brown，S. D.，Francis，R.，Rose，J.，and Vranesic，Z. 1992. *Field-Programmable Gate Arrays.* Kluwer Academic，Boston，MA.

[6] Buchanan，J. 1990. *CMOS/TTL Digital System Design.* McGraw-Hill，New York.

[7] Deyhimy，I. 1985. GaAs digital ICs promise speed and lower cost. *Computer Design* 35（11）：88-92.

[8] Jenkins，J. H. 1994. *Designing with FPGAs and CPLDs.* Prentice-Hall，Englewood Cliffs，NJ.

[9] Kanopoulos，N. 1989. *Gallium Arsenide Digital Integrated Circuits*：A System Perspective. Prentice-Hall，Englewood Cliffs，NJ.

[10] Leigh，B. 1993. Complex PLD &FPGA architectures，*ASIC&EDA.*（Feb.）：44-50.

[11] Matisof B. 1986. *Handbook of Electrostatic Control（ESD）*. Van Nostrand Reinhold，New York.

[12] Matthew，P. 1984. *Choosing and Using ECL.* McGraw-Hill，New York.

[13] Rosenstark，S. 1994. *Transmission Lines in Computer Engneering.* McGraw-Hill，New York.

[14] Scarlett，J. A. 1972. *Transistor-Transistor Logic and Its Interconnections.* Van Nostrand Reinhold，London.

[15] Schilling，D. and Belove，C. 1989. *Electronic Circuits：Discrete and Integrated*，3rd ed. McGraw-Hill，New York.

[16] Texas Instruments，1984. *Hign-Speed CMOS Logic Data Book.* Texas Instruments，Dallas，TX.

[17] Xilinx，1993. *The Programmable Gate Array Data Book.* Xilinx，Inc.，San Jose，CA.

备注

为专业读者考虑，下面列出了关于离散逻辑和可编程器件最新发展动向的宝贵技术期刊资料：

《*Electronic Design*》：由美国俄亥俄州克利夫兰市的Penton出版社发行，该刊物在一年之内刊印了30次，其内容几乎涵盖了所有与电路设计相关的主题，向会员免费开放。

《*EDN*》：由NewYork的Cahners出版公司出版，每年发行38期。该杂志也是几乎涵盖了所有与电路设计相关的主题，并向会员免费开放。

《*Computer Design*》：由美国新罕布什尔州Nashua市的Pennwell出版公司出版，每月一期。该杂志主要是关于数字设计主题，也同样向会员免费开放。

《*Integrated System Design*》：由美国加利福尼亚州洛杉矶市的 Verecom 集团每月发行一期，是关于可编程逻辑器件和相关设计工具当前信息的很有价值的资料。

对于研究人员来说，《*IEEE Journal of Solid State Circuits*》可能是关于数字逻辑器件最新发展的最佳参考资料。IEEE 每年都主办了很多关于数字逻辑和可编程逻辑器件的各种会议。

制造商的数据手册和应用注释提供了很多关于逻辑器件的最新发展信息，尤其是应用方面的信息。但是，他们通常都假设了很多背景知识，在第一次见到时是很难懂的。

下面列出了关于这些背景知识的推荐书籍：

[1] Brown, S. D., Francis, R., Rose, J., and Vranesic, Z. 1992. *Field-Programmable Gate Arrays.* Kluwer Academic, Boston, MA.

[2] Buchanan, J. 1990. *CMOS/TTL Digital System Design.* McGraw-Hill, NewYork.

[3] Jenkins, J. H. 1994. *Design with FPGAs and CPLDs.* Prentice-Hall, Englewood Cliffs, NJ.

[4] Matthew, P. 1984. *Choosing and Using ECL.* Mc Graw-Hill, New York.

[5] Rosenstark, S. 1994. *Transmission Lines in Computer Engneering.* McGraw-Hill, New York.

第8章 存 储 器

Shih-Lien Lu

8.1 引言

存储器是所有计算机系统中最基本的器件之一，用来存储计算指令和数据。逻辑上，存储器可以看做是连续的存储单元集合，每个存储单元对应一个惟一的地址作为标签，并可以存储信息。对存储器的访问是通过向器件提供所需数据的地址来完成的。

存储器可以根据它的功能划分为两种类型：只读存储器（Read Only Memory，ROM）和可读写存储器或随机存取存储器（Random Access Memory，RAM）。而 ROM 又可以分为：大多数时候读取、少数时候写入的存储器和 Flash ROM。RAM 也可以根据存储特征分为两种类型：静态 RAM（Static RAM，SRAM）和动态 RAM（Dynamic RAM，DRAM）。其中，DRAM 每隔一段时间就要进行刷新，以防止由于电荷泄露而导致数据出现差错；而 SRAM 则无需刷新。

SRAM 和 DRAM 都是易失性存储器，也就是说，当失去供电电源后，它们的数据将会丢失。而与易失性存储器对应的非易失性存储器，在失去供电电源后，它们的数据仍然可以保留不变。当前所有的 ROM 器件（包括大多数时候读取、少数时候写入的存储器）都是非易失性存储器。除了少部分特殊存储器外，非易失性存储器的连接方式都是类似的。当一个地址被指定给一个存储器而且一个控制信号被过滤后，存储在指定地址中的信息在一定延迟之后可以被重新得到，这个过程成为“存储性读取”，这个延迟作为从有效地址到数据预备所需的时间，称为“存储性读取访问时间”。类似地，也可以通过执行“存储性写入”操作来将数据存储到存储器中。在写入时，通过触发一个写入控制信号来将数据和地址提交给存储器。用于连接的还有其他控制信号，例如，在芯片的封装图上，大多数存储器都有一个片选（或者芯片使能）引脚。只有当这些引脚使用时，对应的存储器才会处于工作状态。一旦一个地址被提交给了一个芯片，内部地址译码逻辑电路就可以准确定位该地址对应的内容。由于电路结构（该结构用来实现译码逻辑电路）性质的原因，存储器通常在连续读取或写入之前需要进行恢复。因此，连续地址读写之间的时间间隔称为“循环时间”。循环时间通常是存取时间的两倍。在存储器中，还有很多其他的时间要求。这些时间参数在存储

器与计算机处理器之间的接口电路中占有很重要的地位。在很多情况下，一个存储器的时间参数会直接影响计算系统的性能。

某些专用的存储器结构不会遵从一般的通过地址访问的流程。其中应用最频繁的两个是内容可编址的存储器（Content Addressable Memory，CAM）和先进先出（First In First Out，FIFO）存储器。另一种类型的存储器称为"多端口存储器"，该存储器可以接收多路地址并在不同端口产生各种输出结果。还存在一种存储器，可以并行写入，串行读取，例如视频 RAM 和 VDRAM，因为它们通常用作图像显示，稍后我们将对其作详细讨论。

8.2 存储器构造

存储器的构造涉及到很多方面，接下来我们将采取自顶向下的方式进行讨论。

1. 存储器分级体系

存储器的运行速度落后于处理器的运行速度。随着处理器变得越来越快，性能越来越完善，就要求存储器具有越来越大的存储容量，以便与处理器中软件复杂度增加的步调取得一致。图 8-1 给出了著名的摩尔定律，图中描述了中央处理单元（Central Processing Unit，CPU）和存储器容量的指数增长规律。虽然 CPU 的速度随着技术和设计工艺（流水线操作）的发展不断变快，但是由于存储器容量的不断增加，更多的时间花在了对越来越宽的地址进行解码以及对存储在不断收缩的存储单元上的信息进行读取的过程上了。因此，CPU 和存储器之间的速度差距变得越来越大了，图 8-1b 阐释了这一现象。

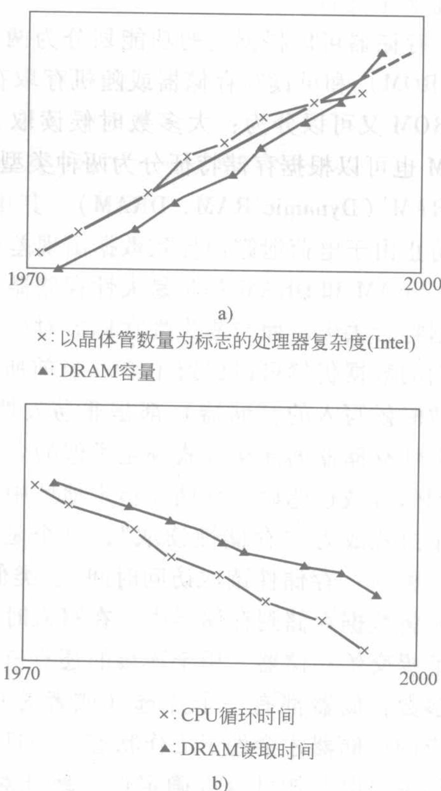


图 8-1　摩尔定律

a）处理器和存储器的发展趋势　b）RAM 和 CPU 之间的速度差距

解决上面这个问题的策略称为"存储器分级体系"。存储器分级体系的基础是存储器参考模型中的位置属性，这些存储器位置属性由连

续读取程序指令和相关数据共同决定。在一个存储器分级体系系统中，包含很多存储器分级。小部分的快速存储器通常紧靠 CPU，以便与存储器和 CPU 的速度相协调。随着存储器和 CPU 之间的距离越来越远，对存储器性能上的要求则不是那么严格了；同时，存储器的容量也变得越来越大，以满足整体存储容量的要求。存储器分级体系包括：寄存器、高速缓冲存储器（Cache）、主存储器和磁盘存储器。图 8-2 给出了常规系统中的一般存储器分级体系。一旦存储器的参考模型确定下来，那么处理器会首先访问分级体系中最顶端的存储器。如果所需的数据位于较高的体系级别中，那么处理器就会得到一个 bit 信息，并可以很快得到要访问的数据；否则，就会产生一个“缺失”信息，所请求的信息必须从体系中较低的级别提升到较高的级别。通常，存储器的存储空间一般被划分成多个存储块，以便在各个级别中转换。在高速缓冲存储器级别上，每个存储区域称为“Cache 块”或“一个 Cache 线”；而在主存储器级别上，每个存储区域称为“存储页”。Cache 中的“缺失”称为“Cache 缺失”；而主存储器上的“缺失”称为“页缺陷”。当出现缺失时，如前所述，包含被请求缺失信息的整个存储器区域就会被从体系中较低的级别提升到较高的级别。如果缺失出现时的存储器分级体系已经是完整的了，那么就必须移出现有的存储块或存储页面，并重新写入到较低级别的存储器中，以便将来再被提升上来。这个过程中包含了几种不同的“替代算法”。其中，最常见的是“最近最少使用（Least Recently Used，LRU）”替代算法。

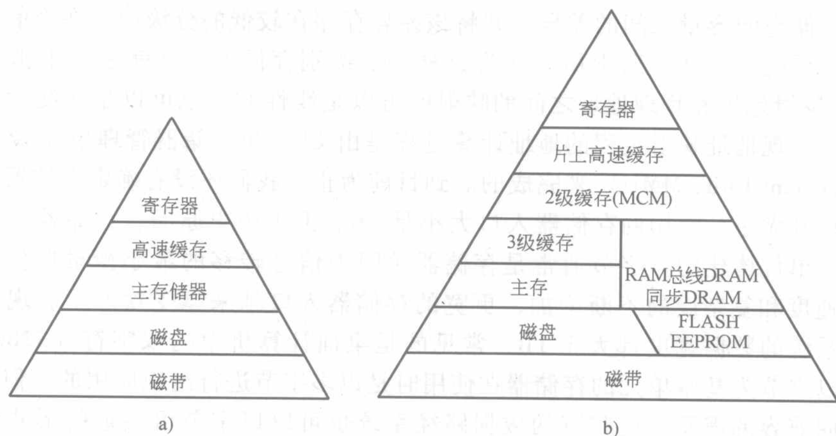


图 8-2　常规分级存储器体系

在现代计算机系统中，Cache 分级中可能包含很多 Cache 子分级。存储器分级体系的基本原理是：离 CPU 越远的存储器，其容量越大，速度越慢，每个存储单元的价格越便宜。由于 CPU 的可编址存储空间通常比专业软件程序所需的

存储空间大，因此，磁盘通常作为主存储器的经济型补充，这种技术称为“虚拟存储器”。除了磁盘外，还有磁带、光学驱动器以及其他后备器件，即我们通常所说的“备用存储器”。备用存储器主要用来存储不再使用的信息，以避免主存储器和磁盘出现故障，或用来在两台机器之间转移数据。

2. 系统级存储器构造

在系统级构造中，我们必须有效组织存储器的结构来完成不同的任务并满足程序的需求。在计算机系统中，地址是由 CPU 提供的，用来访问数据或指令。根据给定的一组地址，一组固定数量的存储器空间就可以被访问了，这些空间就是“存储器地址空间”。有一些处理系统具有访问其他独立空间的能力，这些独立空间称为“输入/输出（I/O）地址空间”；而其他处理系统利用部分的存储器空间作为输入/输出（I/O）的访问地址空间，这种 I/O 功能称为“存储器映射 I/O”。存储器地址空间定义了直接可编址存储器的最大容量，计算机系统可以通过存储器类型指令访问该直接可编址存储器。例如，一个具有 16bit 地址宽度的处理器可以访问高达 64k 个不同的地址（存储器入口地址），而 32bit 地址宽度的处理器则可以访问高达 4G 个不同的地址。但是，有时我们也可以采用间接方式来增加地址空间。例如，80×86 处理器中使用的方法就为这种间接方式给出了很好的解释。存储器系统中用户或程序员用来标识数据项目的地址称为“逻辑地址”。通过逻辑地址访问的地址空间称为“逻辑地址空间”。但是，逻辑地址不必直接用来索引物理存储空间。我们将通过物理地址访问的存储空间称为“物理地址空间”。当逻辑空间比物理空间大时，就需要一个存储器分级体系来调节存储空间容量之间的差异，并将该差异存储在较低的分级中。在当前大部分计算机系统中，硬盘通常用来作为这样的低级别存储器，这就是“虚拟存储系统”。逻辑地址和物理地址之间的映射既可以是线性的，也可以是非线性的。实际中，实现地址映射过程的地址计算过程是由 CPU 和存储器管理单元（Memory Management Unit，MMU）来完成的。到目前为止，我们还没有确定存储器入口宽度的精确大小。常用的存储器入口大小是 1B。由于历史原因，存储器是以字节为基本单位构建的。字节通常是存储器访问中信息转移的最小衡量单位。随着 CPU 速度和复杂度的不断增加，更宽的存储器入口越来越受欢迎了。现在很多先进系统的数据宽度都大于 1B，常见的是桌面计算机中的双字符（32bit）。因此，以字节为基本单元的存储器在使用时是以多字节进行组织应用的。但是，由于后向兼容的需要，这些宽的数据路径系统也可以以字节可编址的形式组织应用。存储器转换的最大宽度称为“存储器字长”，而存储器以字节为单位的存储空间大小称为“存储容量”。由于存储器容量之间存在区别，因此，存储器系统可以根据不同容量的存储器来构建。

3. 存储器的结构

物理上，在一个存储器中，存储单元是呈二维矩阵排列的，每个存储单元可

以存储 1bit 的信息。通过指定所需行和列的地址就可以访问对应的矩阵。每行的使能线通过一个地址译码器生成，每列通过多路复用器来选择。在每列的位线和多路复用器之间通常存在一个读取放大器，用来读取被访问存储单元的数据内容。图 8-3 给出了常见存储单元阵列的示例，图中描述了 $r$bit 的行地址和 $c$bit 的列地址。因此，总的地址位数是 $r+c$，而该存储结构则总共包含了 2^{r+c}位。随着存储阵列数量的不断增加，行使能线和列使能线都会变得更长。为了减小每条行使能线上的电容性负载，行译码器、读取放大器以及列多路复用器通常放置在各个单元阵列之间，如图 8-4 所示。通过对多路复用器进行不同的设计，我们可以

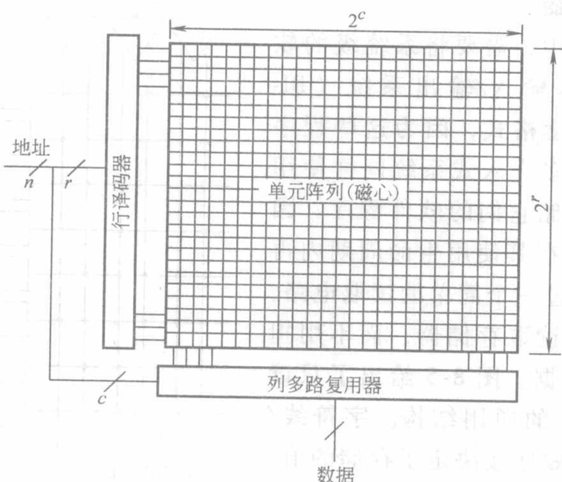


图 8-3　常见的存储件结构

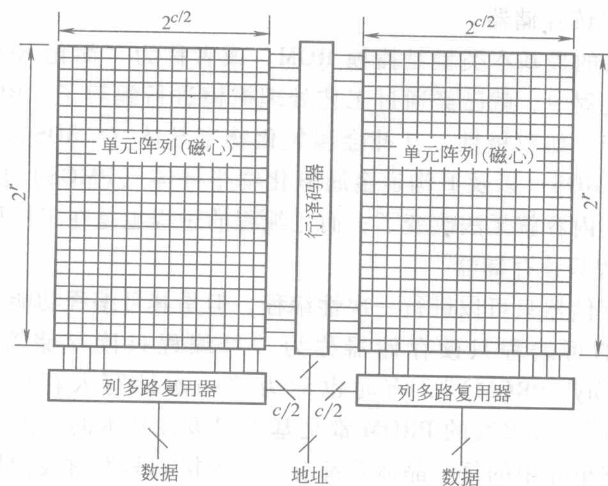


图 8-4　分开的存储器结构

构建具有不同输出宽度的存储器，例如 ×1、×8、×16 等等。事实上，存储器设计人员付出了很多努力来设计列多路复用器，这样大多数的工艺掩模就可以被具有相同容量但不同结构的存储器件共享。

8.3 存储器类型

如前所述，根据存储器的功能和特征，我们将存储器划分为两种主要类型：ROM 和 RAM。在接下来的内容中，我们将详细介绍这两种器件。

1. 只读存储器

在很多系统中，需要将系统级的软件（例如，基本输入/输出系统（BIOS））存储为只读格式，因为这些程序很少被修改。很多嵌入式系统同样使用只读存储器来存储它们的软件程序，因为通常这些程序在其使用生命周期内再不会被改变。通过一个简单的读取电路，就可以很可靠地读取存储器，而不用担心破坏存储的数据。图 8-5 给出了只读存储器（ROM）的通用结构。字符线/位线交叉点的有效连接决定了存储的值。这种连接可以通过不同的技术来实现，从而得到不同类型的 ROM。

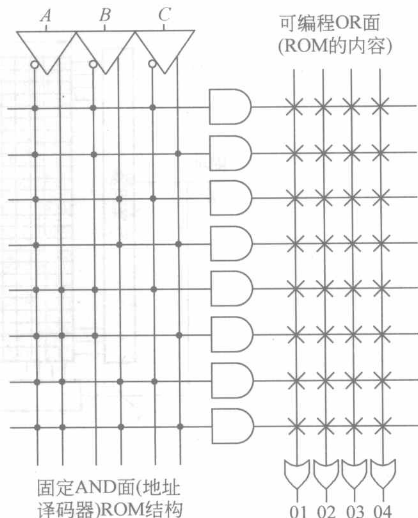


图 8-5 ROM 的通用结构（8×4 ROM）

（1）掩模只读存储器

只读存储器的最基本类型是掩模 ROM，或者称为“简化 ROM”。该类型的存储器在制造过程中，就已经通过工艺处理掩模进行编程了。ROM 可以通过不同的技术来生产，如双极性、互补金属氧化物半导体（CMOS）、N 沟道金属氧化物半导体（NMOS）以及 P 沟道金属氧化物半导体（PMOS）技术等等。一旦它们被编程，其内容就无法改变了；而且编程的过程也是在工厂里面完成的。

（2）可编程只读存储器

有些只读存储器只可以进行一次性编程，但是其可编程功能是由用户来完成的，这种类型的可编程只读存储器称为“可编程只读存储器（Programmable Read-Only Memory，PROM）”，有时也称为“一次性写入存储器（Write Once Memory，WOM）”。大多数的 PROM 都是基于双极性技术的，因为该技术非常适合 PROM。存储单元中的每个晶体管都有一个连接到其发射极的熔丝，晶体管和熔丝构成了该存储单元。当熔丝被熔断后，通过行线选择的存储单元就没有了连

接线，因此一个0值信息就存储在该单元中了。否则，熔丝完整无缺，就代表了一个完整的1值。编程过程是通过程序设计员完成的，该程序设计员称为“PROM程序员”或“PROM熔固员”。图8-6给出了双极性PROM存储单元的结构图以及它在制造时的横截面图。

（3）可擦除可编程只读存储器

有时，只读存储器的一次性可编程特性非常不方便。因此，人们设计出了可擦除的PROM。这种类型的可擦除PROM称为“可擦除可编程只读存储器（EPROM）”。对每个存储单元的编程过程是从衬底穿过氧化层通过雪崩注入高能量电子来实现的。这个过程是通过使用高漏极电压来完成的；在这个过程中，漏极电压使电子获得了足够的能量，从而跃迁过了衬底和二氧化硅之间3.2eV的障碍。另外，漏极电压也在活动的栅极上聚集了电荷。一旦漏极电压消失，这些电荷就被固定在活动的栅极上了。擦除就是通过使用紫外线（UV）擦除器来完成的。UV光增加了固定在活动栅极上电子的能量，一旦电子能量超过3.2eV的障碍，电子就会离开活动栅极并进入衬底和选定的栅极。因此，在这些EPROM芯片的封装上都有一个窗口，以方便UV光照射到芯片的内部，用来擦除存储单元中的内容。擦除的时间通常为数分钟。活动栅极上的电荷会造成MOS晶体管具有一个很高的开启电压。因此，尽管栅极电压为正，该电压应用在多晶硅的第二层，但是MOS晶体管仍然处于截止状态。活动栅极上如果没有电荷，会造成MOS晶体管具有一个很低的开启电压。一旦栅极选定，晶体管将导通并提供相反的数据位。图8-7给出了带有活动栅极的EPROM存储单元的横截面图。朝微型化方向发展的EPROM技术使活动栅极通过UV光照射放电（擦除）变得更加艰难了。其中的一个问题是金属位线的宽度不能随着处理技术的发展而成比例地减小。EPROM金属线的宽度限制了位线的布置，从而减少了到达充电单元的高能光子的数量。因此，基于亚微技术的EPROM器件就需要越来越长的UV线照射时间。

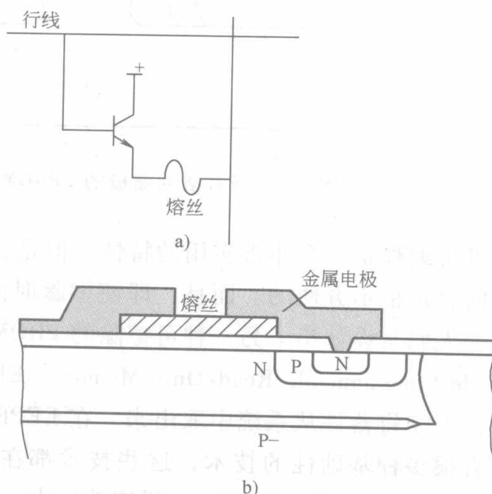


图8-6 双极性PROM
a）双极性PROM存储单元 b）双极性PROM存储单元的横截面

（4）电可擦可编程只读存储器

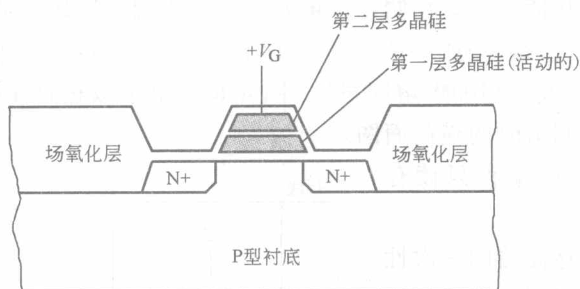


图 8-7 带有活动栅极的 EPROM 存储单元的横截面图

可再编程是一个非常实用的特性，但是，使用独立光源擦除器来修改存储器中的内容是很不方便的。而且，即使擦除时间只有几分钟，但也是无法容忍的。因此，人们又设计出了另一种可擦除的 PROM 称为“电可擦除 PROM（Electrical Erasable Programmable Read-Only Memory，EEPROM）”。EEPROM 在实现编程应用时，无需将器件从系统中取出来。在 EEPROM 或电可再编程 ROM 的处理过程中，有很多种基础性的技术，这些技术都在某种程度上利用了 Fowler-Nordheim 隧道效应。在 Fowler-Nordheim 隧道效应中，能量较低的电子克服了硅-二氧化硅界面的能量障碍并跃迁到了氧化层导带。上面的跃迁过程只有当氧化层厚度只有 100Å 或更小时才可能发生，氧化层厚度与采用的技术有关。Fowler-Nordheim 隧道效应是可逆的，也就是说，允许可再编程 ROM 重复使用。最早的电可擦除 PROM 是电改写只读存储器（Electrical Alterable ROM，EAROM），EAROM 基于金属亚硝酸氧化物硅（Metal Nitrite Oxide Silicon，MNOS）技术。另外一个电可擦除 PROM 就是 EEPROM，EEPROM 基于硅活动栅极技术，该技术应用在 EPROM 的制造工艺中。采用活动栅极技术的 EEPROM 由于可靠性和密度上的优势而更受欢迎。EPROM 和 EEPROM 的主要区别是它们对活动栅极上存储电荷进行放电的方式不同。EEPROM 必须对活动栅极进行电气特性上的放电，与 EPROM 中采用 UV 光源放电的方式大不一样。在 EPROM 中，电子从 UV 辐射出的光子上吸收足够的能量，并克服硅-二氧化硅界面的能量障碍，返回到衬底相反的方向跃迁。EEPROM 的工作原理就是使低能量的电子在高场强（10^7 V/cm^2）的作用下穿过较薄的氧化层，该原理如同“Fowler-Nordheim 隧道效应”一样知名；当场强足够大时，电子就可以穿过很短的绝缘体禁带进入导带。图 8-8 给出了各种不同 EEPROM 的横截面图。1987 年，为了实现小型 EEPROM 存储单元，人们采用了 NAND 结构。在 NAND 结构中，存储单元以串联的形式排列。通过使用不同的样式，我们可以检测出每一个存储单元是否被编程。从用户的角度来说，EEPROM 和 ROM 之间的区别主要在于它们的写入时间和出现差错之前允许

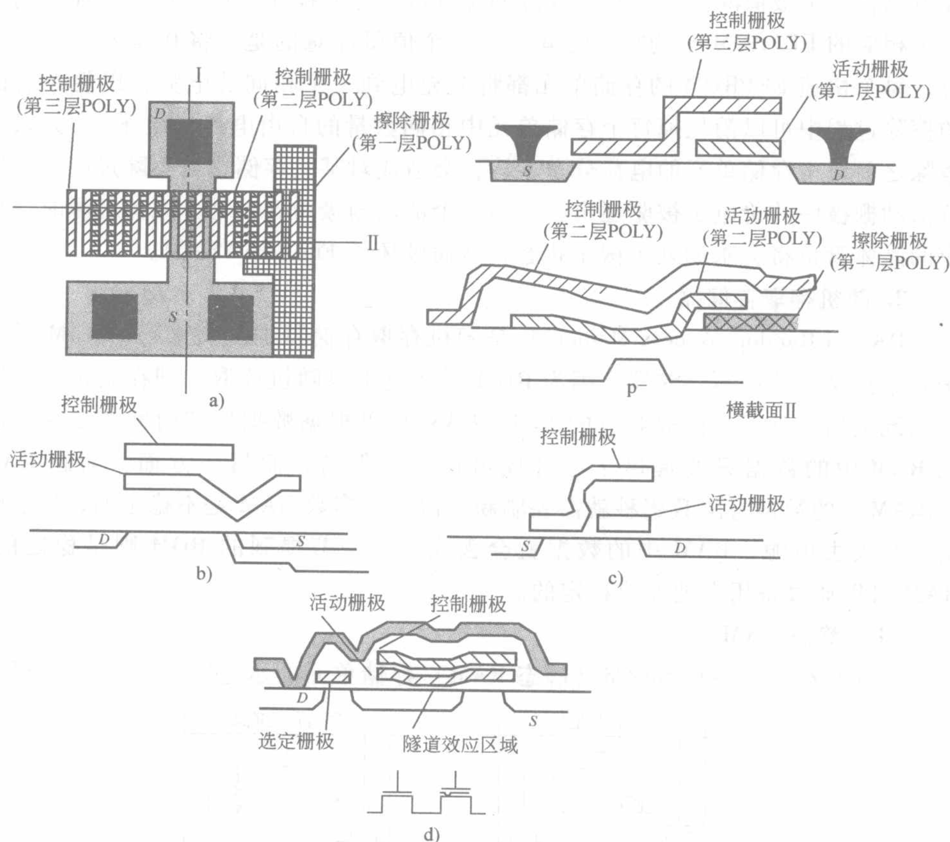


图 8-8 各种 EEPROM 的横截面图

a）三层聚合物 EEPROM 存储单元布局及结构 b）flotox EEPROM（源极编程）
c）漏极编程 EEPROM d）另一个源极编程 EEPROM

写入的次数。早期的 EEPROM 很难使用，因为它在长时间的写入操作过程中没有数据锁存和地址锁存来保持数据，而且还需要一个很高的编程电压，而不是工作电压。较新的 EEPROM 使用了电荷泵来产生芯片上的高编程电压，因此，用户不用提供单独的编程电压。

（5）Flash-EEPROM

Flash-EEPROM 存储单元有 3 种类型。其中一种就是采用可擦除栅极（三层多晶硅），其他两种就是分别采用源极和漏极来擦除。这一类的可擦除 PROM 没有专门用于擦除单个位置数据的电路；当对单个位置进行数据擦除时，就会将整个区域的数据全部擦除掉。这样，就可以省下许多晶体管，而且可以使用更大容量的存储器。但是，值得注意的是，有时在写入之前不一定需要进行擦除操作。

你可以在一个被擦除过但没有写入过的位置进行写入操作，这个过程就涉及到了一个相应的 EEPROM 平均写入时间。另一个值得注意的是，将 0 写入一个位置时，每个快速 EEPROM 的存储单元都将会充电到一个相同的电势，以便在后面的擦除过程中可以消耗掉每个存储单元中相同数量的自由电荷（电子）。如果在擦除之前每个存储单元的电荷分布不均，会造成对某些存储单元擦除过度，并导致活动栅极中的价电子被驱逐出去。当一个活动栅极由于这种方式被耗尽时，相应的晶体管也将会永远处于截止状态，从而破坏了 Flash EEPROM。

2. 随机存取存储器

RAM（Random Access Memory）是随机存取存储器的英文缩写。RAM 是一个真正意义上的读写存储器，因为 ROM 同样也可以随机读取，即在给定一个随机的地址后，就可以读取对应的入口。RAM 可以根据数据保持时间来分类。静态 RAM 中的数据只要提供了电源就可以一直保持；而另一方面，动态 RAM（DRAM）的数据每隔数毫秒就需要刷新一次。大多数 RAM 是不稳定的，这意味着一旦失去电源，RAM 中的数据就会丢失。上一节提到的 ROM 则是稳定的。RAM 可以通过备用电池变成稳定的。

（1）静态 RAM

图 8-9 给出了各种 SRAM（静态 RAM）存储单元的示意图（6T、5T、4T）。

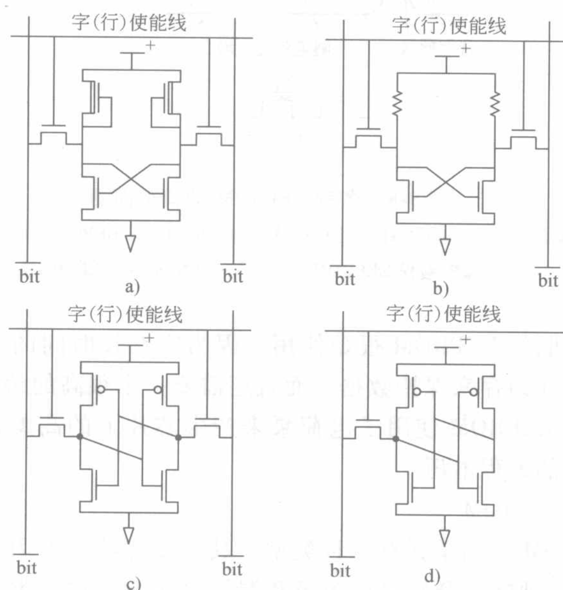


图 8-9　各种不同的 SRAM 存储单元

a）带有耗尽型晶体管负载的 6 晶体管 SRAM 存储单元　b）带有多个电阻的 4 晶体管 SRAM 存储单元　c）CMOS 6 晶体管 SRAM 存储单元　d）5 晶体管 SRAM 存储单元

6 个晶体管（6T）SRAM 存储单元是最常见的 SRAM。只要提供供电电源，SRAM 存储单元中的交叉耦合反相器就可以无限时保持数据信息，因为其中的一个上拉晶体管可以提供电流来补偿泄漏电流。在数据读取的过程中，位和位线被预先充电，同时字节使能线处于低电平状态。当字节使能线被选通时，根据存储单元的内容，其中的一条线会被轻微放电，并造成预先充电的电平下降。这种位和位线之间电压上的差别可以被读取放大器读取，并得到读取结果。在写入的过程中，一条位/位线会被放电，在字节线断开之前，通过选通字节使能线将所需的数据写入存储单元。

图 8-10 给出了一个完整的 SRAM 电路示意图，该电路只有一个行和一个纵列。SRAM 存储单元设计的关键步骤之一就是确定存储单元中晶体管的尺寸，不

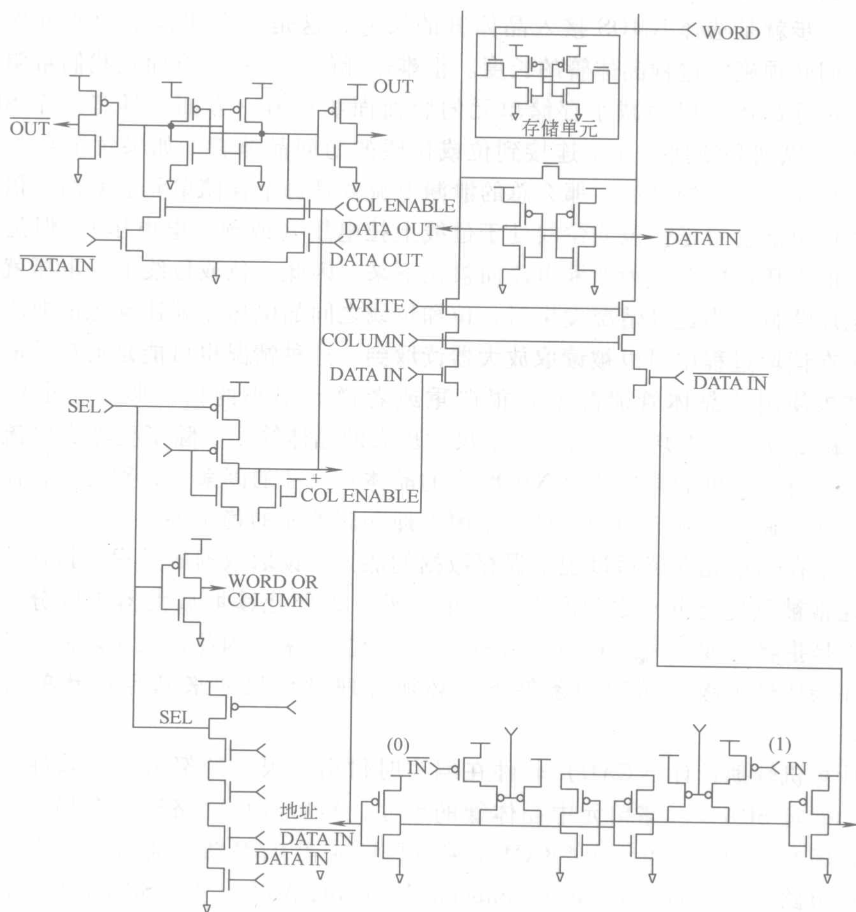


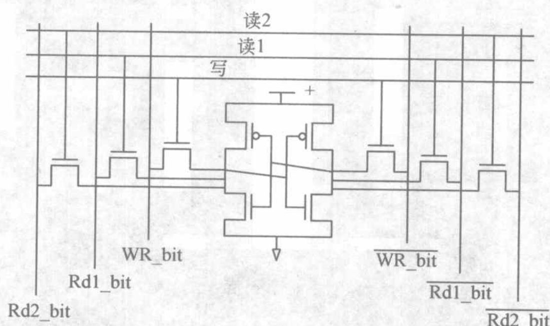
图 8-10 完整的 SRAM 电路示意图

过我们首先必须确定 CMOS 6 晶体管/存储单元 SRAM 中衡量晶体管尺寸的标准。根据对称性质，在 6 晶体管 CMOS SRAM 存储单元中，我们只需要选择 3 种晶体管的尺寸，它们分别是：PMOS 上拉晶体管的尺寸、NMOS 下拉晶体管的尺寸以及 NMOS 访问晶体管（又称为“通过晶体管门”）的尺寸。这也是为什么有些 SRAM 完全摒弃两个 PMOS 晶体管，并采用两个 10GΩ 的多晶硅电阻取代它们的原因，图 8-9 给出了 4 晶体管存储单元的示意图。由于存储单元设计的目标之一就是使存储单元的版面尺寸尽可能得小，因此选择了长度最短和宽度最窄的 PMOS 上拉晶体管。只有当还存在可利用的空间时（例如不增加存储单元的版面面积），PMOS 上拉晶体管的长度才可以增加。通过增加 PMOS 上拉晶体管的长度，可以增加交叉耦合反相器输出端节点处的电容，这可以用来弥补一定的软件错误，也可以使得存储单元的写入操作更容易一些。

下一步就是选择 NMOS 接入晶体管的尺寸，这是一个非常复杂的过程。首先，我们必须确定这种晶体管的长度。很难选择的原因：一方面是我们希望该晶体管的尺寸最小，以便减小存储单元的版面面积；另一方面，具有 n 个 SRAM bit 的每个纵列必须拥有 n 个连接到位或位线的访问晶体管。如果每个存储单元泄漏哪怕是一点点的电流，那么总的泄漏电流就是每个存储单元泄漏的 n 倍。因此，我们可能会考虑使位或位线处于位线上拉电压（或预充电电压），但是位或位线上的电压可能会因为泄漏电流而被拉下来。因此，位或位线上的电压就比预想的电压要低。当这种情况发生时，位和位线之间的电压差就比预想的要小，该电压差在读取过程中可以被读取放大器读取到。这种情况也可能是灾难性的。因此，如果使用的晶体管泄漏不是很严重或者说 n 很小的话，那么最小尺寸的 NMOS 就足够了。否则，就需要一个尺寸更大的晶体管了。除了要考虑电流泄漏之外，还有 3 个可能影响两个 NMOS 中的晶体管尺寸的因素，分别是：存储单元的稳定性、速度、版面面积。第一个因素即存储单元的稳定性，这是一个 DC 现象，是指存储单元在读写过程中保存数据的能力。读取数据的过程是指在位和位线（通常被预先充电）之间产生一个电压差，以便让读取放大器来区分。写入的过程是指将位或位线的其中一条的电压完全拉下来。因此，在实现最大的读写速度并保持最小版面面积的条件下，必须合理设计尺寸来满足存储单元的稳定性。

计算机辅助设计（CAD）软件在编写时付出了很大的努力，该软件可以用来自动确定 SRAM 存储单元中晶体管的尺寸，然后进行电路排版布局。一般来说，排版布局的过程可以由 SRAM 宏编辑程序或者 RAM 编译器完成。在很多专用集成电路（Application Specific Integrated Circuit，ASIC）中，SRAM 模块通常用作传输线输入端。标准的 SRAM 芯片还有很多不同的结构，常见的有 4bit、字和双字（32bit）宽度的 SRAM。

同样，在计算机中还有一种特殊的 SRAM 存储单元，用来实现寄存器功能，称为“多端口存储器”。一般，多端口存储器中的内容可以同时被不同的读取请求读取。图 8-11 给出了一个双读取端口单写入端口 SRAM 存储单元的示例。当对 SRAM 存储单元进行排版时，通常可以与相邻的存储单元共享电源线和接地线。图 8-12 给出了 4 个相邻 SRAM 单元采用普通数字处理设计规则进行排版的示意图，这 4 个存储单元构成的模块可以在一个二维阵列中重复出现，并构成存储器的核心。



8-11　多端口 CMOS SRAM 单元（2 个读取端口，1 个写入端口）

（2）直接随机存取存储器

SRAM 的主要劣势在于其外观尺寸，因为它的每个单独存储单元都由 6 个晶体管构成（或者至少由 4 个晶体管或 2 个电阻构成）。因此，通常使用直接随机存取器（Direct Random Access Memory，DRAM）来提高容量。DRAM 存储单元有各种不同类型的设计，包括：4 晶体管 DRAM 存储单元设计、3 晶体管 DRAM 存储单元设计以及单晶体管 DRAM 存储单元设计。图 8-13 给出了这些存储单元的相应电路图。在 3 晶体管 DRAM 存储单元中，数据写入过程是通过保持 RD 线处于低电压，同时选通总线上所需数据对应的 WR 线来完成的（图 8-13b）。如果需要保存数值 1，当 T2 导通时，T2 的栅极就会充电。T2 上充电的电荷会保留一段时间直到泄漏电流对其放电，使其电压无法导通 T2。当 T2 上还有电荷时，就可以通过对总线进行预充电并选通 RD 线来进行读取的操作。如果数值 1 已经被保存，而 T2 和 T3 此时仍然处于读取数据的阶段，那么会导致总线上的电荷被放掉，但是不断降低的电压可以通过读取放大器重新提升起来（读取放大器是共享的，如图 8-14 所示）。如果数值 0 已经被保存，而总线没有直接接地的路径，那么总线上的电荷会一直被保持。为了进一步减小存储单元的版面面积，通常使用单晶体管存储单元。图 8-13c 给出了带一个电容的单晶体管存储单元示意图。通常，两个纵向排列的存储单元是相互映射、相互对称的，这样有助于进一步减

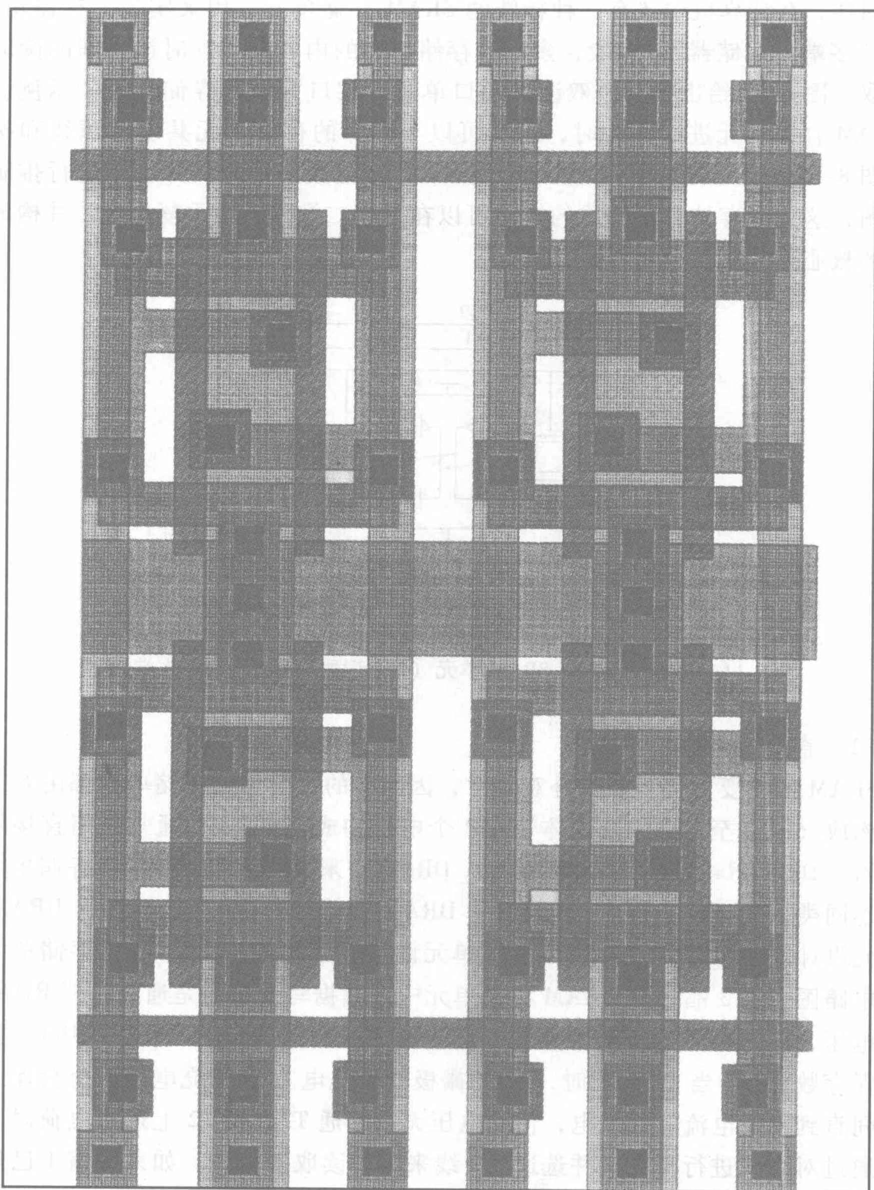


图 8-12　4 个相邻 SRAM 6 晶体管存储单元的排版示意图

小排版时的面积。在这个单晶体管 DRAM 存储单元中，电容是用来存储电荷的，该电荷直接决定了存储的内容；同样，电容中的总电荷决定了存储器的总体性能。下一个目标就是如何减小电容的物理面积来实现更高的集成度。一般来说，随着电容面积的减小，电容的容量也会下降。一种方法就是通过采用堆栈式的电

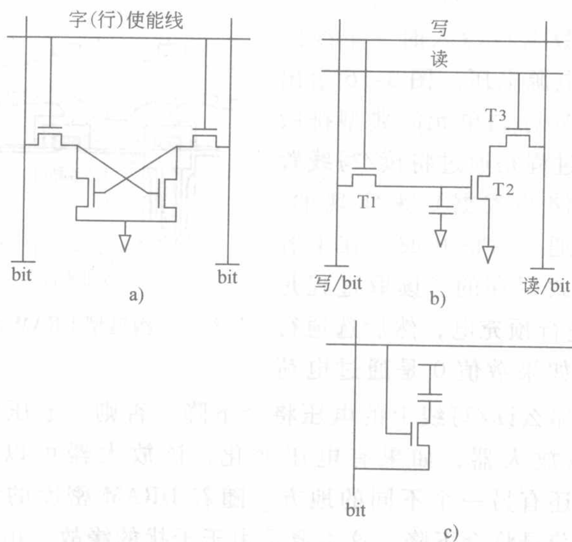


图 8-13 不同类型的 DRAM 存储单元

a）4 晶体管 DRAM 存储单元 b）3 晶体管 DRAM 存储单元

c）单晶体管 DRAM 存储单元

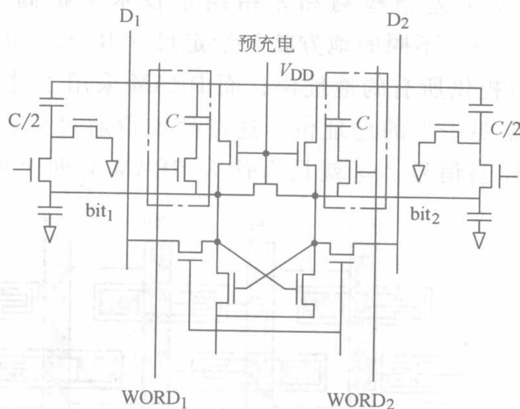


图 8-14 带有 2 条位线和 2 个存储单元的 DRAM 读取放大器

容结构（如垂直结构或圆柱形结构）来增加存储电极的表面面积，同时无需增加排版面积。有些技术可以实现具有半球形晶粒的圆柱形电容结构。图 8-15 给出了具有圆柱形电容结构的单晶体管 DRAM 存储单元的横截面图。由于电容是利用通过晶体管的源极跟随器进行充电的，因此这些电容可以被最大限度地充电直至达到由电源电压产生的开启电压值，从而减小了总体的存储电荷量并影响了性

能、噪声容限和密度。为了避免这个问题的出现，通常数据在写入时，字线的驱动电压要高于电源电压。图8-16给出了单晶体管DRAM存储单元的典型排版示意图。写入的过程是通过将读/写线置0或1（需要存储的数据）来实现的，然后行线就被选通。数值0或1在电容中是以电荷的形式保存的。读取过程是通过对读/写线进行预充电，然后选通行选择来实现的。如果数值0是通过电荷共享来保存的，那么读/写线上的电压将会下降。否则，电压将保持不变。在终端有一个读取放大器，如果有电压变化，该放大器可以用来提升电压。DRAM和SRAM还有另一个不同的地方。随着DRAM密度的增加，每个存储单元中保存的电荷量将会下降，这主要是由于干扰的缘故。由辐射产生的那一类干扰称为"阿尔法粒子"，这些阿尔法粒子都是自然环境中存在的或从DRAM芯片中辐射出来的氦核子。如果阿尔法粒子撞击到了一个存储单元，就有可能改变其存储状态。由于阿尔法粒子只能减少不能消除，有些DRAM制造商就开始想办法制定差错检测和差错纠正技术来提高DRAM的可靠性。DRAM和SRAM还有一个不同的地方就是给定尺寸RAM所需地址引脚的数量。SRAM芯片要求同时提供所有的地址位；而DRAM采用了时分复用地址线，因此在给定时间内只需要一半的地址位。这些地址位通过行和列来划分；同时，还需要一个额外的控制信号。这就是为什么DRAM有两个地址选通信号的原

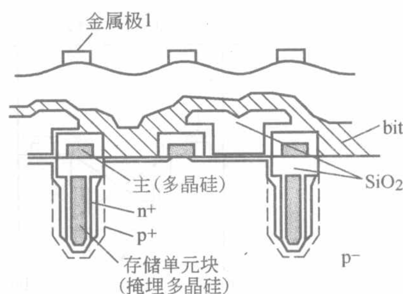


图8-15　沟道型DRAM存储单元的横截面

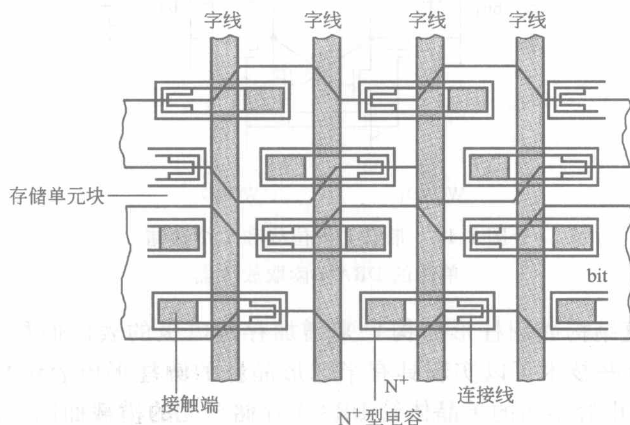


图8-16　沟道型DRAM存储单元的物理排版图

因，这两个选通信号分别是：行地址选通（Row Address Strobe，RAS）信号和列地址选通（Column Address Strobe，CAS）信号。

3. 特殊存储器结构

存储器朝着容量更大、速度更快、性能更好的方向发展，而特殊用途存储器则是朝着互补的方向发展。特殊用途的 RAM 有很多不同的类型，如 Cache 的内容可编址存储器、办公自动化计算机的线性缓冲器（FIFO）、TV 和广播设备的帧缓冲器以及计算机的图形缓冲器等。

（1）内容可编址存储器

存储器中有一类很特殊的存储器，称为“内容可编址存储器（CAM）”或“相联存储器”，该类型的存储器应用在如 Cache 或相联处理器中。CAM 中包含一个数据项，该数据项由一个标签和一个数值构成。提供给 CAM 标签部分的是数据内容，而不是给定的地址；该数据内容和标签部分对应的内容是一致的。如果 CAM 标签部分中的数据项和提供的数据样式匹配，那么 CAM 就会输出与该匹配标签对应的值。图 8-17 给出了 CAM 的基本结构示意图。CAM 存储单元必须既可以读取也可以写入，如同 RAM 一样。图 8-18 给出了基本 CAM 存储单元与匹配标签的电路图，其输出信号可用作某些逻辑器件的输入信号来决定匹配的过程。

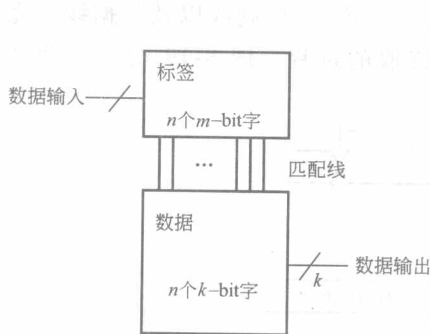


图 8-17 CAM 的功能结构图

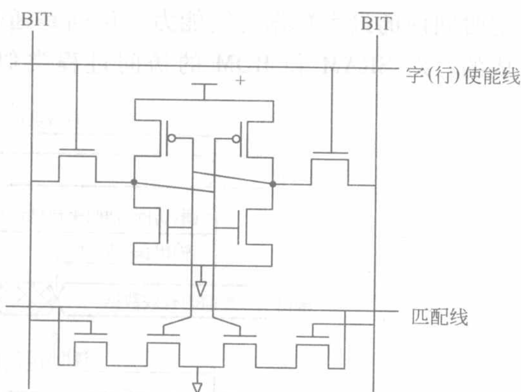


图 8-18 静态 CMOS CAM 存储单元

（2）先进先出/队列

FIFO/队列（First-In First-Out，FIFO/Queue）主要用于等待时的数据保存，如同两个系统的缓冲区域，这两个系统具有不同的数据消耗率和数据生成率。FIFO 可以通过指针移位寄存器或者 RAM 来实现。

（3）视频 RAM：帧缓冲器

计算机图像应用技术正在快速发展，其中最成功的技术称为“光栅扫描”。在光栅扫描显示系统中，每个图像都由一系列的水平线组成，每条水平线都是图像中

相互连接的像素，并控制着图像的亮度。通常，每个主要颜色都对应着3个存储块，主色包括：红色、绿色和蓝色。这3个位图的存储块称为“帧缓冲器”或“图像存储器”。帧缓冲器的结构在很大程度上影响着光栅扫描图像系统的性能。由于这些帧缓冲器在一行一行地显示图像时，需要串行地读出数据，因此，这种特殊类型的DRAM存储器称为“视频存储器”。通常，这些视频存储器具有双端口，其中一个是用作写入时的并行随机访问端口，另一个是用作读取时的串行端口。

8.4 接口存储器

除了容量和类型外，存储器的其他特征参数还包括速度和访问的方式。之前，我们在概述中提到过存储器访问时间，这个时间是指区分地址和访问数据引脚这两个操作之间的时间间隔。有时，访问时间是通过一个特殊的控制信号来测定的。例如，读取控制线准备和数据准备之间的时间称为“读取命令访问时间”。存储器的循环时间是指两个连续访问之间的最小时间间隔；存储器的写入命令时间是指写入控制准备到数据存储进存储器中的时间间隔；存储器的等待时间是指CPU指定一个地址到数据等待处理的时间间隔；存储器的带宽是指在给定时间内的最大数据传输能力。访问是通过地址、读/写控制线以及数据线来完成的。对SRAM和ROM的访问过程类似于读取的过程。图8-19给出了两个

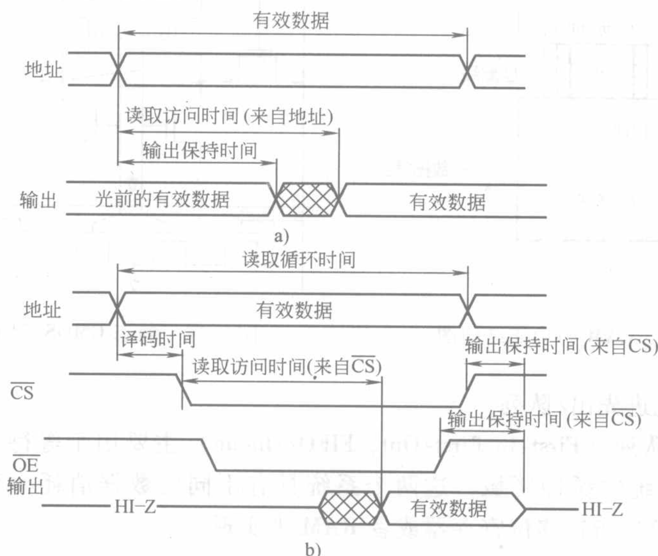


图8-19 SRAM读取过程

a) SRAM简单读取循环（OE~，CE~全部被激活，WE处于低压状态） b) SRAM读取循环

SRAM 读取周期的时间分布图。在这两个过程中，读取循环时间是指连续读取地址之间的时间间隔。在第一个过程中，SRAM 类似于一个异步电路。给定一个地址后，SRAM 的输出值开始变化并在一定延迟之后变得有效，这个延迟时间就是读取访问时间。第二个过程利用了两个控制信号（片选信号和输出使能信号）来触发读取访问过程。这两个过程的主要区别在于数据输出有效时间。在第二个过程中，数据的输出只在输出使能信号触发之后才开始，这使得许多器件都可以连接到数据总线上来。SRAM 的写入过程和电可再编程 ROM 的写入过程有些不同。由于存在很多不同类型的可编程 ROM，而且其写入过程也取决于使用的技术，因此我们在这里将不讨论 ROM 的写入过程。图 8-20 给出了典型 SRAM 芯片的写入过程时间分布图。其中，图 8-20a 给出了采用写入使能信号作为控制信号的写入循环过程；而图 8-20b 给出了采用芯片使能信号的写入循环过程。DRAM 的访问过程和 SRAM、ROM 大不相同。下面，我们将讨论 DRAM 的不同访问模式。

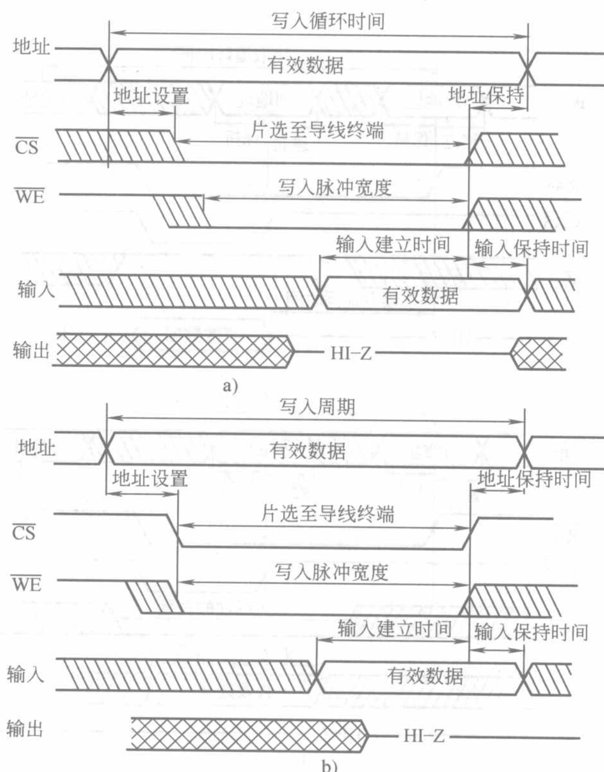


图 8-20　SRAM 写入过程

a）写入使能控制　b）芯片使能控制

1. DRAM 访问过程

DRAM 与 SRAM 最大的不同之处在于 DRAM 的行地址和列地址是时分复用的，这样可以减小芯片引脚的尺寸。由于采用了时分复用，DRAM 的地址就有两条地址选通线，即 RAS 和 CAS。下面列出了 5 种最常见的访问方式。

（1）正常读/写模式

在读取过程中，首先给定一个行地址，其后是行地址选通（Row Address Strode，RAS）信号，RAS 信号用来锁定芯片上的行地址。RAS 信号之后是列地址，列地址之后是列地址选通（Column Address Strode，CAS）信号。在一定延迟（读取访问时间）之后，有效数据就会出现在数据线上面。存储器的写入过程是类似于读取过程，同时只保留读/写控制信号。只有 3 个循环时间可以用来对 DRAM 进行写入，分别是早期写入循环时间、读取修改写入循环时间和后期写入循环时间。图 8-21 给出了 DRAM 芯片的早期写入循环时间示例。另外两种循环时间可以参考 DRAM 的数据手册。

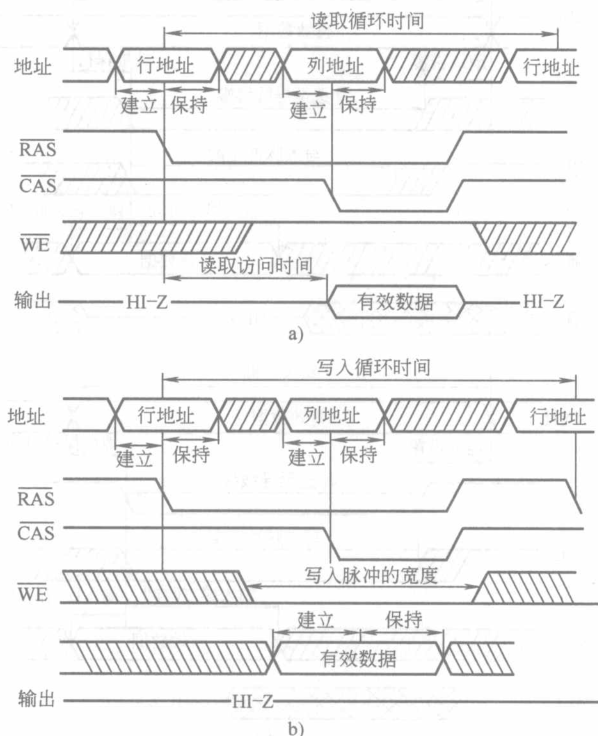


图 8-21　DRAM 读/写循环

a）读取过程　b）写入过程

（2）快速页面模式或页面模式

在页面模式（快速页面模式（Fast Page Mode，FPM））中，读取过程是通过在行地址已知时降低RAM速度来完成的。然后，不管是否有新的行地址，重复提供列地址和CAS信号，而无需循环使用RAS线。在这种模式中，二维阵列（矩阵）的整个行都可以通过一个RAS信号和相同的行地址来访问。图8-22给出了页面模式DRAM芯片上读取时间循环的示例。

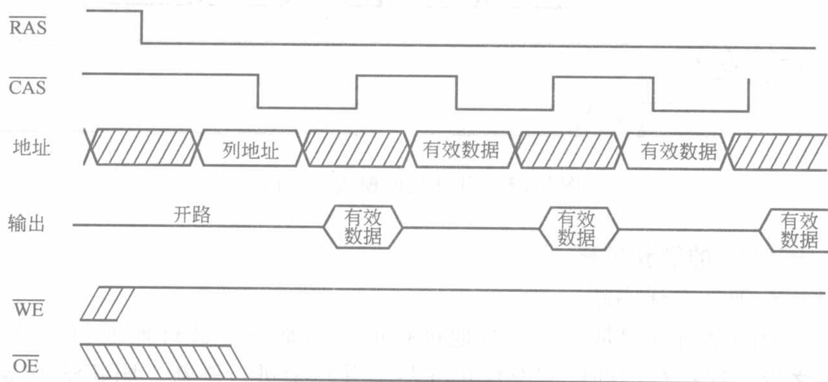


图8-22　页面模式读取时间循环

（3）静态列模式

静态列模式和页面模式类似，不同之处在于当一个新的列地址给定后，CAS信号不是循环的。这个新的列地址就是静态列的名称。

（4）扩展数据输出模式

在页面模式中，CAS必须保持低电压直至有效数据到达输出端。一旦CAS信号触发被取消，数据就会被取消使能，而且输出端引脚就变成了开路。在扩展数据输出（Extended Data Output，EDO）模式的DRAM中，读取放大器后面的附加锁存可以使CAS线的电压迅速回升到高电压值，并可以使存储器开始预充电以备下一次的访问。另外，CAS线回升到高电压之后，数据也会被取消使能。在突发EDO模式的DRAM中，不仅仅CAS线可以回升到高电压，而且该电压值也会被锁定，尽管突发计数器模式中的时序可以在存储器和主处理器之间提供更快的数据传输速度。图8-23给出了EDO页面模式DRAM芯片的读取循环示例。EDO模式也被某些制造商称为“超页面模式（Hyper Page Mode，HPM）”。

（5）半字节模式

在半字节模式中，给定一个列地址的情况下，一个CAS信号之后将会自动执行3次访问，而无需另外的列地址（列地址假设是给定列地址的增加值）。

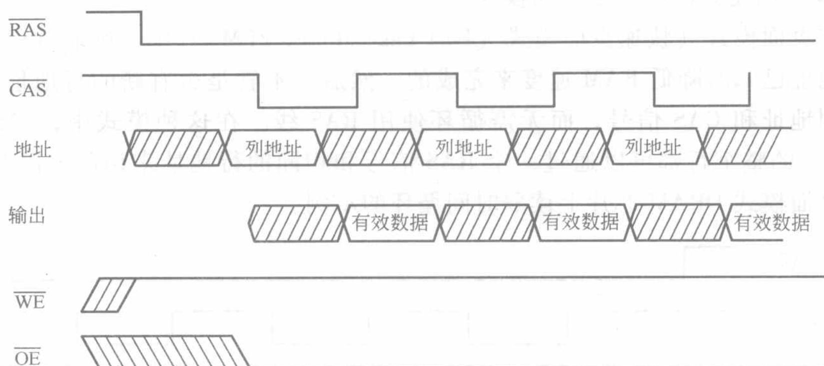


图 8-23 EDO 页面模式读取循环

2. DRAM 的刷新过程

（1）行地址选通刷新

该类型的刷新过程是一行一行地进行的。当某一行被行地址和选通 RAS 信号选中之后，这一行中的所有存储单元将会并行地进行刷新。刷新整个器件耗费的循环周期次数和存储器中行的数量相同。例如，一个 1M×1 的 DRAM，其行列数均为 1024，因此，刷新整个器件将耗费 1024 个循环周期。为了减小刷新循环周期的次数，存储器阵列的排列有时会出现行少列多的情况。但是，地址会被复用成两个均匀划分的字（在 1M×1DRAM 中，每行每列的地址字节宽度是 10bit）。地址线上的高位通常在内部用作列地址线，而且在刷新过程中会被忽略。行地址选通刷新过程无需 CAS 信号。由于 DRAM 的输出缓冲器只有当 CAS 信号触发后才会被使能，因此，在行地址选通刷新过程中，数据总线不会受到影响。

（2）隐蔽刷新

在正常的读取循环时间内，RAS 和 CAS 信号一般都是在提供了相应的行和列地址后才被选通的。某些 RAS 信号可能会在对行地址进行刷新的同时被触发，而在读取完成之后 CAS 信号将不会保存为高电压值。这种刷新样式称为“隐蔽刷新循环”。另外，由于 CAS 信号被选通但没有被保存，因此输出数据不会受到刷新循环过程的影响。刷新循环时间的数量会受到 CAS 信号保持触发最大时间的限制。

（3）RAS 刷新前的列地址选通

为了简化并加快刷新过程，可能会用一个片上刷新计数器来为阵列产生刷新地址。在这种情况下，就需要一个独立的控制引脚来向 DRAM 发送信号，从而触发刷新循环过程。但是，由于在正常的操作过程中，RAS 信号通常是在读写

CAS之前就被触发，因此，可以利用相反的情况来发送刷新循环启动信号。这样，在目前的自动刷新DRAM中，如果控制信号CAS在RAS信号之前被触发，那么它就会发送刷新循环启动信号，我们称之为“RAS之前的CAS刷新”，而且这也是1Mbit DRAM中最常见的刷新模式。但是，还有一个不同的地方需要注意。在这种刷新循环过程中，$\overline{WE}$引脚不会对1Mbit DRAM芯片产生影响。但是，4Mbit DRAM就会通过将$\overline{WE}$引脚电压置高来指定RAS之前的CAS刷新模式。如果$\overline{WE}$引脚处于低电压状态，一个RAS之前的CAS刷新循环将会使4Mbit DRAM进入JEDEC-指定测试模式（WCBR）。相反，通过将测试引脚电压置高可以进入1Mbit DRAM测试模式。

上面所提到的3种刷新循环过程可以通过两种方式在器件上实现。第一种方式是利用一种分布式方法，第二种方式是利用一种等待并突发的方法。采用第一种方式时，器件在常规速率下利用CBR刷新计数器来刷新每行，该计数器每次只启动一行。在这种类型的系统中，只要不是正在刷新过程中，自动刷新过程一完成就可以访问DRAM了。第一个CBR脉冲应该在DRAM使用之前的外部刷新率周期内产生，以确保数据的完整性，而且CBR刷新必须在3个外部刷新率周期的时间内完成。由于CBR刷新通常作为标准刷新来执行，因此，相比第二种方式，这种在触发自动刷新之后就可以访问DRAM的能力就成了实际应用中的一种优势。而第二种方式则是利用外部突发刷新方案来实现的。传感器电路通常用来检测存储单元的电压，来判断它们是否需要更新，而不是在常规间隔时间内启动某行。刷新的过程是通过一系列的刷新循环时间来完成的，该循环时间是一个接一个的，直到刷新完成。在刷新过程中，不允许外部访问DRAM。

8.5　检错与纠错

大多数DRAM都需要一个奇偶校验位，原因有两个：第一，阿尔法粒子的撞击会通过电离辐射扰乱存储单元，并造成数据丢失；第二，在读取DRAM时，存储单元的存储机制会通过使能（选择）晶体管和位线对电荷进行电容性共享。这样会产生一个很小的电压差，该电压差在读取过程中会体现出来，而且会受到其他邻近位线电压和干扰的影响。为了得到更加稳定可靠的存储器数据，必须使用纠错编码。有一种纠错编码方式称为“汉明（纠错）编码”，汉明编码可以纠正任何1bit的错误。

名词解释

动态随机存取存储器（DRAM）：这类存储器是动态的，因为它需要定期进

行刷新。而随机存取是指既可以读取，也可以写入。

存储器访问时间：从有效地址提供给存储器开始到数据在器件输出端准备好之间的时间间隔。

存储器循环时间：存储器中访问连续地址之间的时间间隔。

存储器分级体系：将存储器以分级的形式进行组织，以便使存储器在速度上与处理器匹配。

存储器读取：从存储器中提取数据信息的过程。

存储器写入：将数据信息保存到存储器的过程。

ROM：只读存储器的英文缩写形式。

静态随机存取存储器（SRAM）：这类存储器是静态的，因为它不需要刷新。而随机存取是指既可以读取，也可以写入。

参考文献

[1] Alexandridis, N. 1993. *Design of Microprocess-Based Systems.* Prentice-Hall, Englewood Cliffs, NJ.

[2] Chang, S. S. L. 1980. Multiple-read single-write memory and its applications. *IEEE Trans.* Comp. C-29 (8).

[3] Chou, N. J. et. al. 1972. Effects of insulator thickness fluctuations on MNOS charge storage characteristics. *IEEE Trans. Elec. Dev.* ED-19: 198.

[4] Denning, P. J. 1968. The working set model foe program behavior. CACM 11 (5).

[5] Flannigan, S. and Chappell, B. 1986. J. *Solid St. Cir.*

[6] Fukuma, M. et. al. 1993. Memory LSI reliability. *Proc. IEEE* 81 (5). May.

[7] Hamming, R. W. 1950. Error detecting and error correcting codes. *Bell Syst. J.* 29 (April).

[8] Katz, R. H. et. al. 1989. Disk system architectures for high performance computing. *Proc. IEEE* 77 (12).

[9] Lundstrom, K. I. and Svensson, C. M. 1972. Properties of MNOS structures. *IEEE Trans. Elec. Dev.* ED-19: 826.

[10] Masuoka, F. et. al. 1984. A new flash EEPROM cell using triple poly-silicon technology. *IEEE Tech. Dig.* IEDM: 464-467.

[11] Micro. *Micro* DRAM Databook.

[12] Mukherjee, S. et. al. 1985. A single transistor EEPROM cell and its implementation in a 512 K CMOS EEPROM. *IEEE Tech. Dig.* IEDM: 616-619.

[13] NEC. n. d. *NEC Memory Product Databook.*

[14] Pohm, A. V. and Agrawal, O. P. 1983. *Hign-Speed Memory Systems.* Reston Pub., Reston, VA.

[15] Prince, B. and Gunnar Due-Gundersen, G. 1983. *Semiconductor Memory.* Wiley, New York.

[16] Ross, E. C. and Wallmark, J. T. 1969. Theory of the switching behavior of MIS memory transistors. *RCA Rev.* 30: 366.

[17] Samachisa, G. et al. 1987. A 128 K flash EEPROM using double poly-silicon technology. *IEEE International Solid State Circuit Conference*, 76-77.

[18] Scheibe, A. and Schulte, H. 1977. Technology of a new n-channel one-transistor EAROM cell called SIMOS. *IEEE Trans. Elec. Dev.* ED-24 (5).

[19] Seiichi Aritome, et al. 1993. Reliability issues of flash memory cells. *Proc.* IEEE 81 (5).

[20] Shoji, M. 1989. *CMOS Digital Circuit Technology*. Prentice-Hall, Englewood Cliffs, NJ.
[21] Slater, M. 1989. *Design of Microprocessor-bases Systems*.

备注

关于存储器构造和分级体系基本要素的信息，读者可以参考 Pohm 和 Agrawal 在 1983 年出版的著作。另外，Prince 和 Due-Gunderson 在 1983 年出版的著作也提供了很好的关于不同存储器的背景知识。关于更新的存储器技术，读者可以参考存储器数据手册，如 Mukherjee 等人在 1985 年出版的著作和 NEC 数据手册等。*IEEE Journal on Solid-State Circuits* 每年都会出版一期关于国际固态电路会议要点的资料，这个会议探讨了当前存储器的快速发展情况，如 DRAM、SRAM、EEPROM 以及 flash ROM。关于存储器技术的要点，读者可以参考《*IEEE Transactions on Electron Devices*》。以上两本期刊都一个年鉴索引，该索引通常在每年年底（12 月份）出版。

第9章　微处理器

James G. Cottle

9.1　引言

在通常情况下，我们认为微处理器就是指芯片上的中央处理单元（CPU）。微处理器的技术优势发展得非常快，其发展的动力来自于极大规模集成（Ultra Large Scale Integration，ULSI）电路/超大规模集成（VLSI）电路的物理学性质和技术、工艺的先进性（包括尺寸的变小）以及结构上的改进。

基于微处理器的系统开发人员关心的是成本、性能、功耗以及可编程的难度。其中，后面的一点可能是将产品迅速带入市场最重要的因素。可编程难度中的关键项目是程序开发工具的可用性，因为这一点可以节省小型系统实现过程中的大量时间，也可以使开发人员的开发过程变得更加容易。我们可以很容易发现很多小型电子系统是由微控制器和微处理器来驱动的，该微处理器的功能比小型电子系统所需的简单功能复杂得多，因为这些平台上的开发容易程度抵消了成本和功耗上的顾虑。因此，采用微处理器之后，开发工具就成了系统有效实现过程中非常重要的因素。

通常，常用的微处理器还有一个相同系列的微控制器作辅助。微控制器的主体是大面积嵌入式系统的一个子集。微控制器包含了父微处理器的基本结构，该父微处理器上具有附加的片上专用硬件，用来制造较容易的小型系统。这些专用硬件包括模数转换（Analog/Digital，A/D）电路、数模转换（D/A）电路、定时器、少量的片上存储器、并行和串行接口以及其他输入输出指定硬件。这些器件对小型电子系统的开发非常有用，因为所有外部需要的硬件都已经集成到了芯片上，因此，微控制器之外的元器件数量通常可以保持最少。另外，父微处理器的先进开发工具也可用于编程。因此，一个具有发展潜力的产品推向市场的速度就比采用常用微处理器开发出来的产品要快得多，采用常用微处理器开发出来的产品通常需要很多外部硬件支持。

9.2　体系结构的基本要素

在早期的大型计算机时代，只有很少一些指令可供程序员使用，而且 CPU

还必须与各种应用相适应，或者改变 CPU 的配置来与各种应用相匹配。由 Maurice Wilkes 在 1949 年发明的微程序设计技术释放了指令集。微程序设计可以通过明确定义但可编程的方式合理利用 CPU 的资源（CPU 的寄存器、算术逻辑单元和内部总线）来执行更复杂的任务。在这种类型的 CPU 中，微程序设计或控制存储器会抑制 CPU 内部资源中一个特殊子集的主机指令语义的实现，但是微程序设计的应用目前仍然很广泛。微程序设计中包含了用来实现更复杂指令和编址模式的程序，这就是为什么微程序设计在目前仍然应用很广泛而且是复杂指令集计算机微处理器关键要素的原因。

从最原始的采用电线作逻辑路径和应用指令操作方案设计的大型计算机，向更加灵活和更加复杂的指令集和编址模式的方向发展，这是一个很自然的发展趋势。其发展动力来自于半导体制造工艺、存储器设计以及元器件物理学的发展。半导体制造工艺的发展使得将 CPU 放置在一个单块集成电路上成为了可能，同时还配有附加的其他硬件。随着可编程只读存储器的发明，微程序设计就不必受到原始设计者想法的限制了；而且日后，如果一个特殊任务的资源配置需要进行改进或简化时就可以进行再编程。从最简单的角度来看，随着指令集和编址模式越来越多，微处理器变得越来越复杂了。所有这些发展都受到了编译器开发者的欢迎，而且指令集的复杂度是元器件可靠性和集成度发展的自然结果。

1. 复杂指令集计算机和精简指令系统计算机处理器

微处理器的性能是根据高速条件下执行简单逻辑任务的能力来评估的。在复杂指令集计算机（Complex Instruction Set Computer，CISC）微处理器中，小型寄存器组、存储器对存储器的操作、大型指令集（具有不同的指令长度）以及微指令的使用是最常见的。CISC 微处理器的简化基本原理是指添加的硬件可提高整体的速度。次末级的 CISC 微处理器中的每条高级语言语句都和单个本地 CPU 的指令相互映射。微指令在某种程度上简化了复杂度，但是必须采用多个计算周期来执行单个 CISC 指令。在 CISC 微处理器对指令进行编译之后，实际执行过程只需要 10 个或 12 个计算周期了，具体多少个周期取决于所使用的指令和编址模式。微处理器最初的趋势是朝着指令集复杂度不断增加的方向发展，虽然有成百上千条本地机器指令，但是实际中用到的只有一少部分。具有讽刺意味的是，CISC 的指令集复杂度的发展是以牺牲速度为代价的，因为很难提高复杂芯片的时钟速度。最近，由于精简指令系统计算机（Reduced Instruction Set Computer，RISC）的发展，提高时钟速度又受到了人们的关注。

在 20 世纪 70 年代，反向跟踪研究的趋势表明，尽管 CISC 处理器具有大量的指令，但是只有相对小的一部分指令才会在实际应用中被程序设计人员使用到。事实上，85% 的程序都只包含简单的赋值指令（如 A = B）。

RISC 处理器使用很少的指令和计算周期来实现各种功能，而且它只有一些

寄存器和并行接口。理想 RISC 处理器能够在单个计算周期内执行一条完整的指令，因此，一个 100MHz 的微处理器应该能够在每秒钟内执行一百万条指令。很多用于转移数据的寄存器可以用来缩短计算周期。因此，RISC 处理器就具有很高的并行度。换句话说，CISC 处理器拥有相对较少的寄存器，具体数量取决于数据微程序的多个计算周期处理过程。在 CISC 处理器中，微程序可以处理本地机器指令解释当中的很多复杂问题。这种先进的微处理器通常拥有超过 100 条的本地机器指令和很多的编址模式。

以上这两个基本处理器体系（RISC 和 CISC）都是理想的概念，都有各自的优点。在实际应用中，微处理器合并了这两种体系的原理设计方案，以此来提高处理器的性能和速度。关于 RISC 与 CISC 的比较概括如下文所述。由于 CISC 处理器的性能取决于微程序的编程复杂度，从而可以对编译器进行简化；而 RISC 处理器复杂度则是直接通过编译器自身来实现的。

2. 逻辑相似处

微处理器通常专用于某个特定需求，但是，所有微处理器的外部逻辑特征都是相同的。虽然器件的封装是有区别的，但是它包含了各种控制信号的连接引脚，这些控制信号要么来自于微处理器，要么就是送入到微处理器的。这些引脚连接到系统的供电电源、外部逻辑以及其他硬件，以此构成完整的计算系统。这些硬件包括外部存储器、算术处理器、时钟发生器以及中断控制器，所有这些硬件都有一个与一般微处理器的接口。图 9-1 给出了典型微处理器的引脚引线和常见控制信号组的示意图；图 9-2 给出了微处理器 Intel 8086 和 8088 的结构示意

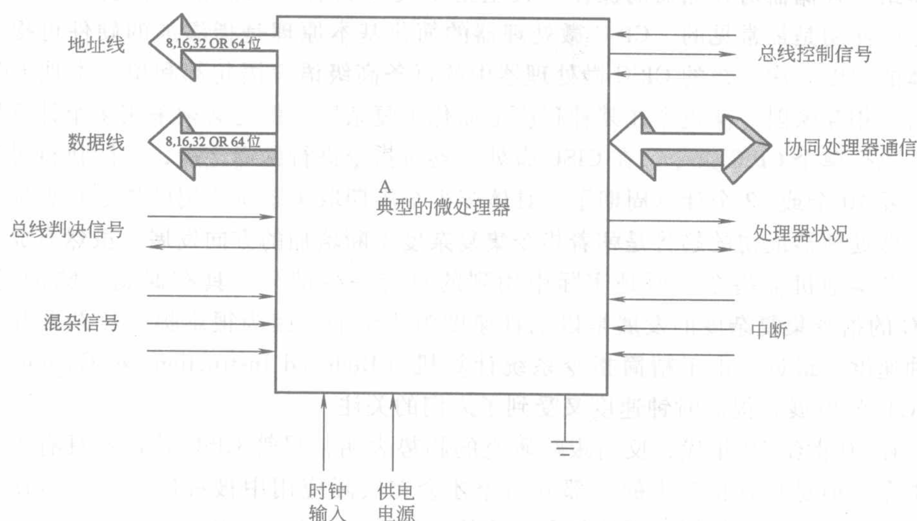


图 9-1　典型微处理器的引脚引线和常见控制信号组示意图

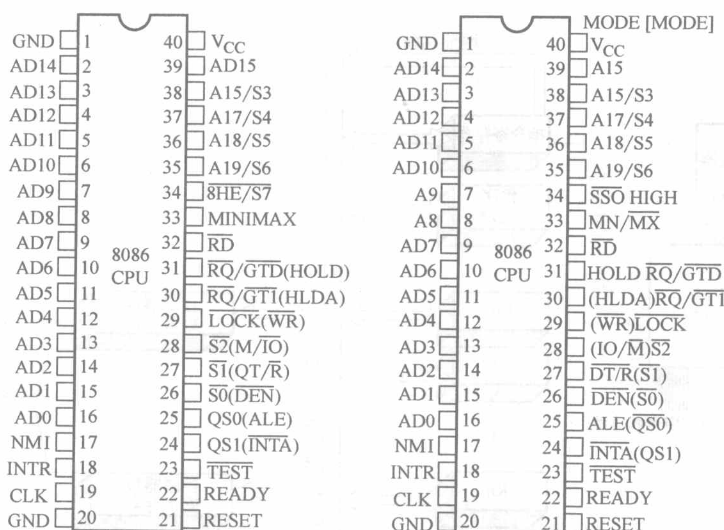


图 9-2 Intel 8086 和 8088 微处理器的引脚引线示意图

图。外部信号包括编址信号、数据信号、总线判决和控制信号、协同处理器信号、中断信号、状态信号以及混杂的连接信号。在某些情况下，多个引脚可用于连接，如地址和数据总线。

图 9-1 所有微处理器连接到外部硬件的引脚都是相同的。它们包括编址引脚、数据总线判决和控制引脚、协同处理器和中断引脚以及连接到电源和始终发生器芯片的引脚。

由于各个系列微处理器的逻辑连接都是相同的，因此其基本指令也是相同的。例如，所有处理器都有一个加法指令、一个减法指令以及一个存储器读取指令。尽管逻辑上这些指令都是相同的，但是各个处理器执行指令（如加法指令 ADD）语义的方式不一定相同。ADD 指令的执行方式取决于 CPU 的内部资源，如可用寄存器的数量、内部总线的数量以及是否在相同的内部传输路径上将数据和地址信息分离。

为了更好理解微程序是如何执行 CISC 处理器中本地机器语言指令的语义的，我们可以参考图 9-3。图 9-3 解释了单总线简单 CISC 处理器的原理，这条总线用来传输微处理器内部的存储信息和数据信息。在这种情况下，CISC 处理器同时只能执行一条指令，因此就比目前大多数的微处理器简单多了。但是，它还是阐释了指令执行的过程。这个过程由一个程序（一个控制存储的程序，称为“取指令”）开始，该程序从存储器中提取出下一条连续指令，并将其内容放到指令寄存器中用于解释。而程序计数寄存器中包含了这条指令的地址，其内容放置在存储器地址寄存器中，以便在稍后存储器读取命令执行过程中可以使下一条指令的内容出现在存储

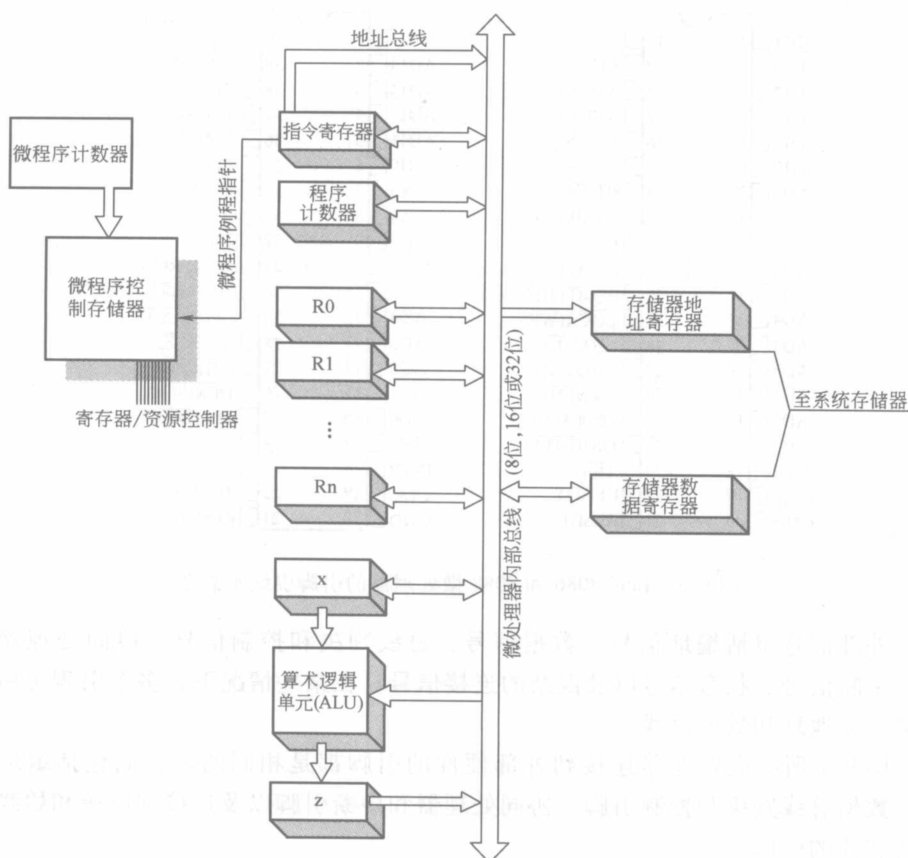


图 9-3 基本单总线计算机结构

器数据寄存器当中。然后这些内容将会沿着总线传输到指令寄存器中以待解码。同时，程序计数器就会增加计数，并包含了下一条指令的地址。指令寄存器中当前指令的内容包含了操作码和关于操作数的地址信息，这些操作数与编码操作相关。例如，如果取出的指令是加法指令，指令寄存器中将包含标示加法的编码位以及用来获取加法操作数的相关信息。处理器的指令位数是不变的（如 8 位，16 位，32 位的指令），但是用于操作码和操作字段的位数会发生变化以适应单操作数或多操作数指令、不同的编码方式等。操作码通常是一个编码地址，该编码地址和微处理器微程序中的位置有关。对于加法指令来说，操作码指示了微程序存储器中的某个位置，该存储器包含了执行加法指令的逻辑代码。加法指令需要许多计算周期，其中包括提取指令自身步骤所需的计算周期。这些步骤和具体的处理器有关。也就是说，所有的处理器都包含有加法指令。微程序中包含了在特定处理器结构上执行加法指令的过程，该结构具有独立的硬件资源。

这个 CISC 处理器由多个寄存器来表征，这些寄存器由总线连接，总线用来承载来自于寄存器或送到寄存器的地址和数据信息。控制线负责每个寄存器的输入和输出端的选通，并由控制存储器进行管理。

引导微处理器从存储器中提取下一条指令的指定步骤（微操作）中还伴随有本地指令。指令提取是每条本地指令的主要部分，代表了最少的计算周期，即微处理器必须在这个计算周期内执行最简单的指令。因此，即使是停机指令也需要几个计算周期来实现。

CISC 处理器方案的优势是很明显的。在大量的可利用指令条件下，编译器或高级语言解释程序编程人员的任务就变得相对简单多了，而且手边可利用的工具也有很多。这些工具包括提高相关编址效率的编址方式方案、直接编址以及自增型编址或自减型编址方式。另外，还有很多存储器之间数据传输指令。但是，编程难度的减小也会付出一定的代价。CISC 处理器方案的劣势在于可能会耗费过多的计算周期，从而使运行速度变得相对较慢。同样，CISC 处理器的结构方案还需要半导体芯片上的许多资源来支持。因此，编程难度的减小会以牺牲尺寸和速度为代价。

Von Neumann 结构曾经是大多数 CISC 处理器的基础。这种结构的主要特点是通用数据和地址总线连接了少量的寄存器。Von Neumann 处理器中寄存器的数量取决于具体设计，但是一般都包含大约 10 个寄存器，其中包括用于保存常见程序数据和地址的寄存器以及用于保存专用程序数据和地址的寄存器，这些专用寄存器包括指令存储器、存储器编址寄存器以及算术逻辑单元（Arithmetic Logical Unit，ALU）的锁存寄存器等。图 9-3 给出了 Von Neumann 处理器的基本结构。但是，最新的发展已经使 CISC 处理器从基本单总线系统中脱离出来了。

3. 微处理器的短期发展历程

第一个单芯片 CPU 是 Intel 4004，由计算器开发而来。它以 4bit 为单位处理数据，而其指令长度是 8bit，其程序和数据是相互分离的。4004 具有 46 条指令、1 个 4 级堆栈、1 个 12 位的程序计数器和 16 个 4 位寄存器。稍后，在 1972 年，4004 的继承者 8008 被发明，其后是 1974 年发明的 8080。8080 具有 1 个 16 位的地址总线、1 个 8 位的数据总线、7 个 8 位的寄存器，1 个 16 位的堆栈存储器指针以及 1 个 16 位的程序计数器。8080 还有 256 个输入输出（I/O）引脚，这样 I/O 器件就不会占用存储器的空间，而且可以更加直接地进行编址。8080 的设计在 1976 年得到了更新，更新后只需要提供 +5V 的电源电压。

Zilog 公司在 1976 年 7 月推出了 Z-80，Z-80 是改进型的 8080。它同样使用 8 位的数据和 16 位的地址，而且可以执行 8080 的所有代码，并添加了 80 多条指令。Z-80 中寄存器组的数量翻倍了，由两组构成，这两组可以相互转换。两个索引寄存器（IX 和 IY）支持更加复杂的存储器指令。Z-80 最成功的地方可能是

它的存储器接口。动态随机访问存储器（RAM）需要相当复杂的外部电路来进行定时刷新，这使得小型计算系统变得更加复杂和昂贵，这种情况直到Z-80诞生才发生了改变。Z-80是第一个将这种片上刷新功能集成到内部的芯片，这就使其在系统开发商中变得非常受欢迎。Z-8是一个嵌入式处理器，类似于Z-80，具有片上RAM和只读存储器（ROM）。Z-8可工作于20MHz的时钟频率，可用于各种小型的基于微处理器的控制系统。

微处理器发展历程中的下一个代表是1975年的6800。6800由Motorola公司发明，而且通过其650x系列，使MOS工艺逐渐成了主流。其中最主要的是6502，该处理器用于早期的桌面计算机（Commodors、Apples和Ataris）。6502具有很少的寄存器，原则上是一个8位处理器，但它具有一个16位的地址总线。Apple Ⅱ作为第一个推向主流消费者市场的计算机，它合并了6502。后来Apple微处理器生产线也开始朝着向下兼容6502的方向改进。1977年，Motorola公司推出了6502的扩展系列6809，该处理器具有两个8位的计算器，该计算器可以将数学运算合并成单个16位的运算，另外，6809具有59条指令。6800系列的主要成员都是以嵌入式微控制器（如68HC05、68HC11）为辅助基础。这些微控制器现在仍然在很多小型控制系统中很受欢迎。68HC11可以扩展成16位，称为“68HC16”。由于68HC11具有抗辐射的功能，因此已经被应用到了通信卫星当中。

后来，美国的AMD公司发明了4位位片微处理器Am2901。位片处理器是标准器件，它可以组合在一起形成更大的字节长度。Am2901具有一个4位的ALU、16个4位的寄存器以及用来连接相邻模块进位/借位信号的硬件。1979年，AMD公司为微处理器开发了第一个浮点协同处理器。AMD 9511算术电路用于某些CP/M、基于Z-80的系统以及某些基于S-100总线的系统。

到1976年，16位微处理器的市场竞争已经达到了白热化。Texas Instruments（TI）公司的TM 9900系列是第一个真正意义上的16位微处理器，并被设计成了单芯片的TI 990小型计算机。TM 9900具有两个16位寄存器、良好的中断处理能力以及为编译器开发人员设计的很好的指令集。TI公司还开发出了嵌入式版本的微处理器（TMS 9940）。1976年，IBM选择将该微处理器用于个人计算机IBM-PC生产线的生产，从而开创了历史。那时的16位微处理器都具有很高的功耗，而且具有很简单的开放式存储器结构（如Motorola公司的68000系列）。当时曾谣传，IBM自己的工程师想使用68000系列，但IBM已经将68000系列的专利权卖给了Intel公司的8086。很显然，选择8位的8088是一个很经济的决定，因为8088使用更加便宜的和8085相关的支持芯片，而68000器件更加昂贵，而且不容易使用。

1976年，Zilog公司推出了Z-8000系列（稍晚于Intel公司的8086）。Z-8000

是一个16位的微处理器，该处理器具有高达23位的地址数据编址能力。Z-8000具有16个16位寄存器。前8个寄存器可以作为16个8位的寄存器使用，或者所有的16个寄存器都可以作为8个32位的寄存器使用，这就为编程和算术运算提供了很高的灵活性。Z-8000的指令集包括一个32位的乘法和除法指令。同样，类似于Z-80，Z-8000具有片内存储器刷新电路。但是，在CPU的发展历程当中，最重要的可能是，Z-8000是第一个将两种不同的运算模式合并在一起的微处理器。其中一个模式被操作系统严格保留；另一个模式被一般用户使用。这种设计提高了稳定性，因为用户不可能轻易地破坏系统；而且还为芯片朝多任务、多用户操作系统（如UNIX）方向的发展开辟了道路。

4. 微处理器的Intel系列

1971年，Intel公司首先采用4004微处理器在芯片上开发出了CPU。这个芯片配合8008可以实现计算器和终端控制的功能。Intel公司并没有对这个单元寄予太多的期望，稍后就开发出了用途更广泛的微处理器8080，以及与8080类似且具有更多板上硬件的芯片8085。这是工业中第一个真正意义上的通用CPU，可以集成到微计算系统中去。Intel公司又在1978年开发出了第一个16位的芯片8086，该芯片是第一个进入工业领域的16位处理器。与8086配合使用的协同处理器8087可用来满足比8086中16位寄存器更高精度的数学计算要求。在开发出8086和16位地址/8位数据版本的8088之后不久，IBM公司选择将8088作为其IBM PC的微型计算机。这个决定为Intel公司带来了巨大的利润，也为其微处理器带来了巨大的发展。在某些方面，Intel公司的早期成功也付出了代价，即让80×86系列成为了牺牲品。因此，后来所有的改进和向更大的数据和地址总线CPU方向的发展都必须考虑后向的兼容性问题。

通常来说，80186和80188微处理器是8086和8088的改进型，它们合并了更多的片上输入和输出支持硬件。但是，80186和80188从来没有得到广泛的应用，这主要是因为被8088在IBM PC上的成功光芒掩盖了。80186在结构上和8086是相同的，但同时它还包含1个时钟发生器、1个可编程控制中断、3个16位可编程定时器、2个可编程DMA控制器、1个片选单元、可编程控制寄存器、1个总线接口单元以及1个6B的预取堆栈。

从20世纪80年代到现在，Intel系列中没有一个处理器具有对1MB的存储器进行编址的能力。80286作为一个具有68个引脚的微处理器，主要用来满足系统和程序发展的需要，这主要得益于8088的成功。80286增加了Intel微处理器系列的可用地址空间，存储器容量可以达到16MB。同样，从80286开始，连接到芯片外部的数据和地址线不再是共享的了。在早期的芯片中，地址引脚线是和数据线复用的。内部结构在实现这一点时有些繁琐，但是内部结构还必须保持与早期CPU的后向兼容性。除了笨拙一点之外，80286还是比较成功的。

在十年的时间里，微处理器已经从最早的起步阶段（4位的CPU）发展到了真正的16位微处理器。许多日常的计算任务已经从大型计算机转移到了桌面计算机上。1985年，Intel公司开发出了真正的片上32位处理器80386。80386兼容所有后向处理器直到8008的目标代码，并继续将Intel系列锁定在笨拙的存储器模型（80286开发包含的）中。在Motorola公司方面，68000在某些方面具有更加简单和直接的开放式地址空间，成为了大型计算应用（从大型计算机到桌面计算机）的重要竞争者。正是由于这个原因，即使在今天我们仍然可以在需要与UNIX操作系统兼容的系统中发现68000。不过，80386还是非常成功的。80386SX是80386的一个系列，具有和80286相同的封装，是现有80286系统的一个更新版本。其更新之处是将80286的地址更新为32位，但仍然保持16位的数据总线。数学协同处理器（浮点数学单元（Float Point math Unit，FPU））80387用来配合80386使用。

80386的成功提醒了其他半导体公司（尤其是AMD和Cyrix公司）通过克隆处理器来为终端用户和系统开发商提供可替代的资源。随着Intel公司在1989年推出80486，处理器芯片之间的竞争就更加激烈了。80486中包含完整的总线、片上高速缓存以及集成的浮点处理器（不是独立的浮点处理器）。在1993年下半年，Intel公司不再延用下一个连贯的名称了（80586）。Intel注册了“Pentium”作为80586处理器的商标。由于它的流行性，80x86生产线被广泛进行克隆。

5. 微处理器的Motorola系列

与Intel公司开发的8080齐名的是Motorola公司开发的6800系列，该系列是一个8位的微处理器。在20世纪70年代早期，6800通常用于许多嵌入式的工业控制系统。但是，直到1979年，Motorola公司才推出它的16位处理器68000，以此来进入工业领域。68000的设计在很多方面都远比Intel公司的8086微处理器先进。其所有的内部寄存器都是32位宽度，而且可以对16MB容量的外部存储器进行编址，而无需Intel系列中的分段设计方案。无需分段意味着68000不需要分段寄存器，而且每条指令都可以对存储器的完整补码进行编址。

68000是第一个合并了32位内部寄存器的16位微处理器。这一点使得设计师可以将混杂操作系统统一到桌面计算机上来。在某些方面，68000是它所处时代的领跑者。如果IBM公司选择68000系列作为其个人计算机的核心芯片，那么现在的艺术级桌面计算机将会从根本上有所不同。Apple公司选择了68000作为Macintosh型号计算机的处理器芯片。其他计算机厂商，包括Amiga和Atari都因为其灵活性和大容量的内部寄存器选择了68000。

1982年，Motorola公司向市场推出了另一款芯片68008，68008是68000的一个简化版本，它是一款低档和低成本的产品。68008只能对4MB容量的存储

器进行编址，而且其数据总线只有 8 位宽。因此，68008 从来没有流行过，也无法和 Intel 公司的 8088 系列（被 IBM 公司选作 PC 计算机的芯片）进行竞争。

先进操作系统对于 68000 来说是非常重要的，除非芯片不能支持虚拟内存。由于这个原因，Motorola 公司开发了 68010，68010 可以在由于总线出现错误而延缓的情况下继续执行指令。68012 和 68000 几乎一样，除了 68012 具有 30 个地址总线引线并可以对 2GB 容量的存储器进行编址之外。

Motorola 公司推出的最成功的微处理器之一是 68020。68020 是在 1984 年推出的，它是工业中第一个真正意义上的 32 位微处理器。68020 遵循 68000 系列的 32 位寄存器标准，并可以对 4GB 容量的存储器进行编址，而且具有真正的 32 位宽的数据总线。68020 在现在仍然应用广泛。

68020 包含一个外部 256B 的 Cache 存储器。这是一个指令 Cache，最高可支持 64 条长字节类型的指令。该 Cache 不允许直接访问，仅仅作为一个先进的预取堆栈来使能 68020 执行指令的紧密循环并伴随取指令操作。由于取指令操作占用了处理的时间，因此 68020 中的 256B 指令 Cache 可以在很大程度上提高处理器的速度。

Cache 的性能在 68030 中得到了扩展，它包含了一个 256B 的数据 Cache。另外，68030 还包含一个板上分页存储器管理单元（Paged Memory Management Unit，PMMU），用来控制对虚拟内存的访问。这就是 68020 和 68030 之间的主要区别。PMMU 在 68020 中可以作为一个外部芯片（68851）存在，但是它和 68030 处于同一块芯片上。68030 也包含一个改进的总线接口设计。从外形上看，68020 和 68030 的连接几乎是相同的。68030 可工作于两个时钟频率，分别为：MC68030RC16 工作于 16MHz 的时钟频率，MC68030RC20 工作于 20MHz 的时钟频率。

68000 设计了超级用户和一般用户模式。其设计还可扩展，并可以在一条指令的执行过程中提取下一条指令（这就是 2 阶段流水线技术）。68040 具有 4 阶段流水线技术。680×0 系列的发展一直持续到 1994 年的 68060。68060 是一个超标量体系结构微处理器，类似于 Intel 公司的 Pentium 系列，真正代表了 CISC 和 RISC 体系结构的合并。68060 的 10 阶段流水线技术将 680×0 的指令转换成一个译码的、RISC 类似的形式，并利用资源重命名方案来重新对指令的执行进行排序。68060 具有省电特征，因此在 3.3V 的电源电压情况下，68060 就会产生中断并停止工作（不同于 Intel 公司的 Pentium 系列）。

6. RISC 处理器的发展

RISC 处理器发展的主要推动者是美国伯克利的加利福尼亚大学和斯坦福大学。Sun Microsystem 公司为他们的高速工作站开发出了 RISC 处理器的伯克利版本［可缩放处理机体系结构（Scalable Processor Architecture，SPARC）］。但是，

这还不是第一个 RISC 处理器。比它更早的有 MIP2000（基于斯坦福大学的设计）、Hewlett Packard PA-RISC CPU 和 AMD29000。

AMD29000 是一个 RISC 设计，紧跟在伯克利设计之后。它具有很多寄存器组，可以分为局部寄存器组和全局寄存器组。在意识到很多 CISC 复杂指令一般情况下都不会使用之后，才开发了具有 64 个全局寄存器的精简指令集处理器。

名词解释

Cache：少量的快速存储器，在物理位置上紧靠 CPU，用来存储处理器立即需要的数据块。Cache 属于存储器分级体系结构。在体系结构中，第一级（L1）Cache 是最小但是最快的；如果 L1 出现错误，访问过程就会转到第二级（L2）Cache，L2 容量更大但速度相对较慢；如果 L2 出现错误，访问过程就会转到主存储器（L3 Cache，如果存在的话）。

流水线技术：一种微结构技术，这种技术将指令的执行过程划分为几个连续的阶段。流水线 CPU 可以允许多条指令同时执行，但是处于不同的计算周期。或者，在数据传输之前就需要发送一个地址。

超标量体系结构：在给定时钟周期内可以执行多条指令的能力。例如，Pentium 处理器具有两条执行流水线（U 和 V），所以它是超标量体系结构的第 2 级。而 Pentium 的协同处理器在每个时钟周期内可以派送和收回 3 条指令，因此，称其为超标量体系结构的第 3 级。

参考文献

[1] The Alpha 21164A：Continued performance leadership. 1995. *Microprocessor Forum.*
[2] Internal architecture of the Alpha 21164 Microprocessor. 1995. *CompCon* 95.
[3] A 300 MHz quad-issue CMOS RISC microprocessor（21164）. 1995. In *ISSC* 95，pp. 182-183.
[4] A 200MHz 64b dual-issue CMOS microprocessor（21064）. 1992. In *ISSC* 92，pp. 106-107.
[5] Hobbit：A high performance，low-power microprocessor. 1993. *CompCon* 93，pp. 88-95.
[6] MIPS R10000 superscalar microprocessor. 1995. *Hot Chips* Ⅶ.
[7] The impact of dynamic execution techniques on the data path design of the P6 processor. 1995. *Hot Chips* Ⅶ.
[8] A 0. 6μm BiCMOS processor with dynamic execution（P6）. 1995. In *ISSC* 95，pp. 176-177.
[9] A 3. 3v0. 6μm BiCMOS superscalar processor（Pentium）. 1994. In *ISSC* 94，pp. 202-203.
[10] An overview of the Intel Pentium processor. 1993. In *CompCon* 93，pp. 60-62.
[11] Superscalar architecture of the P5-x86 next generation processor. 1992. *Hot Chips* Ⅳ.
[12] A 93MHz x86 microprocessor with on-chip L2 cache controller（N586）. 1995. In *ISSC* 95，pp. 172-173.
[13] The AMD K5 processor. 1995. *Hot Chips* Ⅶ.
[14] The PowerPC620 microprocessor：A high performance superscalar RISC microprocessor. *CompCon* 95.
[15] A new PowerPC microprocessor for the low power computing marker（602）. 1995. *CompCon* 95.

[16] 133MHz 64b four-issue CMOS microprocessor (620). In *ISSC* 95, pp. 174-175.

[17] The PowerPC 604 RISC microprocessor. 1994. *IEEE Micro.* (Oct.).

[18] The PowerPC user instruction set architecture. 1994. *IEEE Micro.* (Oct.).

[19] PowerPC 604. 1994. *Hot Chips* Ⅵ.

[20] The PowerPC 603 microprocessor: A low power design for portable applications. 1994. In *ISSC* 94, pp. 307-315.

[21] A 3.0W 75SPECint92 85SPECfp92 superscalar RISC microprocessor (603). 1994. In *ISSC* 94, pp. 212-214.

[22] 601 powerPC microprocessor. 1993. *Hot Chips* Ⅴ.

[23] The PowerPC 601 microprocessor. 1993. In *CompCon* 93, pp. 109-116.

[24] The great dark cloud falls: IBM's choice. In Great Microprocessors of the Past and Present (on-line) Sec. 3. http://www. cpu info. berkeley. edu.

备注

在快速发展的处理器领域，任何一款产品都可能随着时间的流逝而被淘汰。事实上，市场调查表明，随着每一款产品的诞生，这种产品从一种新技术变成一种陈旧技术的时间越来越短。因此，在设计和开发时就需要更加精确以避免产生错误，例如1994年，Intel公司就曾经在他们的Pentium处理器的浮点数学计算机上经历过这种情况。对于小型系统开发商或微处理器用户来说，最重要的是不断推出新产品和性能更好的芯片。幸运的是，现在有一种方法可以解决这个问题。

在加利福尼亚大学的CPU信息中心，伯克利在万维网（WWW）上提供了非常好的关于最新微处理器和微控制器的编译、它们的结构以及具体细节的信息。网站上还包含了芯片尺寸和引脚引线以及制成表的关于微处理器性能和结构的信息，还有参考文献。该网站会定期更新，可以作为小型系统开发商了解这方面信息的重要资源。本章中包含的一些信息也可以在该网站上找到，这些信息是得到允许才引用的。网站地址为http://www. cpu info. berkeley. edu，读者可以在该网站上查询最新的信息。

第10章　D/A和A/D转换器

Susan A. Garrod

10.1　引言

数/模（Digital-to-Analog，D/A）转换过程是指将数字代码转换成连续变化的模拟信号，而模/数（Analog-to-Digital，A/D）转换过程则是一个相反的过程，是指将连续变化的模拟信号转换成数字代码。这种转换过程是实际模拟系统与数字系统接口连接中必需的，因为实际环境监控的都是连续变化的模拟信号，而数字系统存储、解析以及处理的都是模拟信号对应的数字数值。

D/A和A/D转换电路最早是从军用产品发展起来的，如今已经广泛应用于军用产品和消费类产品。在20世纪80年代中期，军用产品决定了很多D/A和A/D转换器件的设计。军用产品在密封性能、抗辐射、振动测试以及军用规范和记录保存等方面都有很高的要求；而成本是很少考虑的，其产品的“低功耗”功率通常达到2.8W。D/A和A/D转换电路主要应用于军用雷达警戒和导航系统、数字示波器、医学成像、红外线系统以及专业视频等领域。

如今，采用D/A和A/D转换电路的产品性能标准已经和早期的不同了。尤其是，当广泛应用于电池供电的消费类产品后，低功耗和高速应用就不断推动着D/A和A/D电路向前发展。

10.2　D/A和A/D转换电路

D/A和A/D转换电路已经被很多制造商制造成了集成电路（IC）。一个很大的IC阵列不仅可以包含D/A和A/D转换电路，还可以包含与其密切相关的电路，例如采样和保持放大器、模拟多路复用器、电压-频率和频率-电压转换器、参考电压、校准器、运算放大器、隔离放大器、设备放大器以及数据采集系统等。IC厂商的数据手册中包含了很多关于这些器件的信息以及应用电路，对设计工程师会有帮助。

本章将要讨论的IC只包括D/A和A/D转换电路。IC中一般要么执行D/A转换，要么执行A/D转换。不过，串行接口IC既可以执行D/A转换，也可以执行A/D转换，主要应用于数字信号处理。

1. D/A 和 A/D 转换电路的性能标准

决定电路性能质量的主要因素是分辨率、采样率、速度以及线性特性。

D/A 转换电路的分辨率是指输出模拟信号的最小变化量；而在 A/D 转换电路中，分辨率则是指可以被系统检测到的最小电压变化量，该变化量会影响最后生成的数字代码。分辨率决定了数字代码的总数或者说是量化级数，该数字代码由电路产生或识别。

D/A 或 A/D 转换电路的分辨率通常会在数字代码位中或系统的最低有效位（Least Significant Bit，LSB）中被指定。一个 n 位的代码可以支持 2^n 个量化级数，或者 2^n-1 个量化级差。随着位数的增加，各个量化级差就会逐渐减小，从而通过模拟信号和数字信号之间的转换就可以进一步提高系统的精确度。系统的分辨率同样可以通过电压的量化级差数来表示。对于 A/D 转换电路来说，分辨率是指可以被系统检测到的最小输入电压。

D/A 或 A/D 转换电路的速度由其执行转换过程所用的时间决定。对于 D/A 转换电路来说，速度由下降时间来表征；而对于 A/D 转换电路来说，速度由转换时间来表征。D/A 转换电路的下降时间随电源电压和数字代码的转换而变化，因此，下降时间可以根据所处的不同情况列成一个数据表格。

A/D 转换电路具有一个最大采样率，用来限制连续转换的速度。采样率是指每秒内模拟信号被采样并转换成数字代码的点数。对于合适的 A/D 转换来说，最小的采样率必须至少是模拟信号最高频率的两倍，这样才能满足奈奎斯特（Nyquist）采样标准。在确定 A/D 转换器的最大采样率时，必须将转换速度和其他时间因素考虑在内。Nyquist A/D 转换器采用的采样率是稍稍高于模拟信号最高频率的两倍。超采样 A/D 转换器采用的采样率是模拟信号最高频率的 N 倍，N 的范围为 2～64。

D/A 和 A/D 转换电路都需要一个参考电压来实现绝对转换精度。有些转换 IC 中有内部参考电压，而其他则需要接受外部的参考电压。对于高性能系统来说，就需要一个外部的精确参考电压，来确保长期的稳定性、负载调整率以及对温度起伏的控制。IC 外部的精确参考电压可以在制造商的数据手册中找到。

测量精确度由转换器的线性特征表征。完整线性特征是指对整个转换范围内线性特征的量度标准，通常定义为相对终点到零值（或偏移量）之间直线的偏离量。完整线性特征也称为“相对精度”。偏移量用来确定转换范围中零值或中点值的参考电压。微分线性特征是指代码转换之间的线性特征，它是转换器单调性的量度标准。如果输入值增加会直接导致输出值增加，那么这种转换器就称为具有“单调性”。

一个转换器的精确度和线性特征值由数据表中代码的 LSB 来指定。线性特征会随着温度变化而变化，因此它的值通常会包含“+25℃”的温度字样，器

件在整个温度范围的值也是这样表示的。

2. D/A 转换过程

数字代码在转换成模拟电压时，将数字代码中的每位指定一个电压分量，然后对整个代码的所有电压分量进行求和。一个常用的 D/A 转换器由精密电阻网络、输入转换电路以及电压转移电路构成，用来将数字代码转换成模拟电流或电压。产生模拟输出电流的 D/A IC 通常比产生模拟输出电压的 D/A IC 具有更短的下降时间以及更好的线性特征。当输出为电流时，设计人员可以通过选择合适的输出放大器将电流转换成电压，以满足指定应用中响应速度的需要。

D/A 转换器一般都有一个固定的或有效的参考电压。参考电压决定了精确转换电路的转换门限电压，精确转换电路用来构建一个可控制的阻抗网络，该阻抗网络反过来可以控制输出信号的值。具有固定参考电压的 D/A 转换器产生一个与数字输入成正比的输出信号。而乘法 D/A 转换器产生的输出信号与可变参考电压和数字代码的乘积成正比。

D/A 转换器可以产生双极性信号：正信号或负信号。四相乘法转换器支持参考信号和二进制代码的值为正或负；同时，四相乘法转换器也可以产生双极性输出信号。

3. D/A 转换器 IC

大多数 D/A 转换器都是为一般用途的控制电路设计的。但是，有些 D/A 转换器是为专用应用设计的，例如视频或图像输出、高清视频显示、超高速信号处理、数字视频录影、数字衰减器以及高速信号发生器。

D/A 转换器 IC 通常具有一些专用的功能，这些功能使它们可以很容易地和微处理器或其他系统连接。微处理器的控制输入端、输入锁存器、缓冲器、输入寄存器以及与标准逻辑系列的兼容性就是 D/A 转换器 IC 经常使用的专用功能。另外，D/A 转换器 IC 通常具有激光微调精度电阻，用来满足用户实现满标性能的需求。

4. A/D 转换过程

模拟信号可以通过多种方法转换成数字代码，这些方法包括积分法、逐次近似计算法、并行（快速）转换、δ 增量调制、脉冲编码调制和 σ 增量调制转换。常见的两种 A/D 转换过程为逐次近似 A/D 转换和并行或快速 A/D 转换。而超高清数字音频或视频系统则需要专用的 A/D 转换技术，这些专用的 A/D 转换技术通常合并了常见的 A/D 转换技术，以及专用的 A/D 转换技术。专用的 A/D 转换技术包括脉冲编码调制（Pulse Code Modulation，PCM）和 σ 增量调制转换。PCM 是常见的音频编码方案，不仅应用于音频领域的数字录音，还应用于通信产业的音频编码和多路复用。σ 增量调制转换是一种超采样 A/D 转换，其信号的采样频率非常高，而且具有很高的分辨率和较低的失真。

逐次近似 A/D 转换是一种在中高速数据采集应用中非常常见的技术，也是最快的 A/D 转换技术之一，这种技术需要的电路数量最少。8 位系统的逐次近似 A/D 转换过程的转换时间为 10～300μs。

逐次近似 A/D 转换器可以通过 n 个近似步骤将模拟信号转换成一个 n 位的数字代码。逐次近似寄存器（Successive Approximation Register，SAR）会逐个将模拟输入电压与 n 级量化范围中的某一个量化范围的中间值进行比较，来决定其数字代码值是否为 1。这个过程一直重复 n 次，涉及 n 个量化范围，以此来决定数字代码中的 n 数据位。其比较过程如下所述：SAR 判断模拟输入电压值是高于还是低于中间值，然后相应设定数字代码中的数据位。SAR 负责分配数据位，第一位是最高有效位（Most Significant Bit，MSB）。如果模拟输入电压值高于中间值，那么相应的数据位就设置为 1；如果模拟输入电压值低于中间值，那么相应的数据位就设置为 0。然后 SAR 就转移到下一个数据位，并根据模拟输入与下一个量化范围中间值的比较结果来设置该数据位为 1 或 0。由于 SAR 对数字代码中的每一数据位都必须执行一次近似过程，因此，一个 n 位的数字代码就需要执行 n 次近似过程。

逐次近似 A/D 转换器由 4 个功能模块组成，如图 10-1 所示，分别为 SAR、模拟比较器、D/A 转换器和时钟电路。

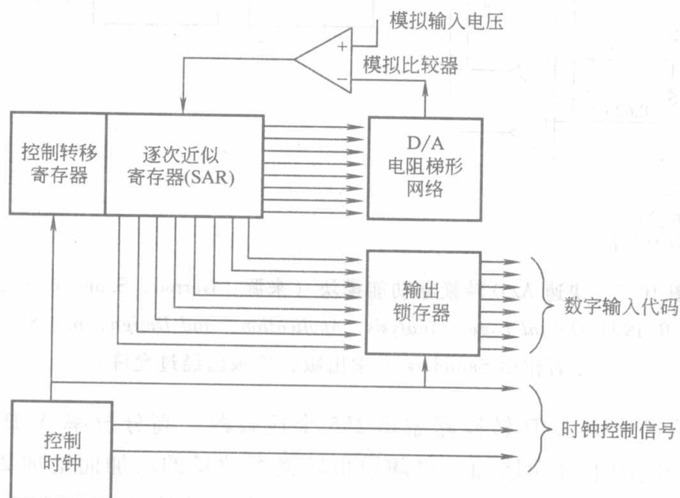


图 10-1 逐次近似 A/D 转换器功能模块（来源：Garrod，S. and Borns，R. 1991. *Digital Logic*：*Analysis*，*Application*，*and Design*，p. 919. 该著作由 Saunders 大学出版，再版已经过允许）

并行或快速 A/D 转换常见于高速应用中，如视频信号处理、医学成像以及雷达探测系统。快速 A/D 转换器同时将输入模拟电压与 2^n-1 个门限电压进行

比较，产生一个 n 位的数字代码来代表模拟电压。分辨率为 8 位的典型快速 A/D 转换器的工作频率为 20～100MHz。

图 10-2 给出了快速 A/D 转换器的功能模块。快速 A/D 转换器电路由 1 个精确梯形网络、2^n-1 个模拟比较器和 1 个数字优先编码器组成。电阻网络用来确定每个量化等级的门限电压；模拟比较器用来指示每个门限电压上的输入模拟电压是高于门限电压还是低于门限电压。模拟比较器的输出信号作为数字优先编码器的输入信号。优先编码器产生最后的数字输出代码，该数字代码保存在输出锁存器中。

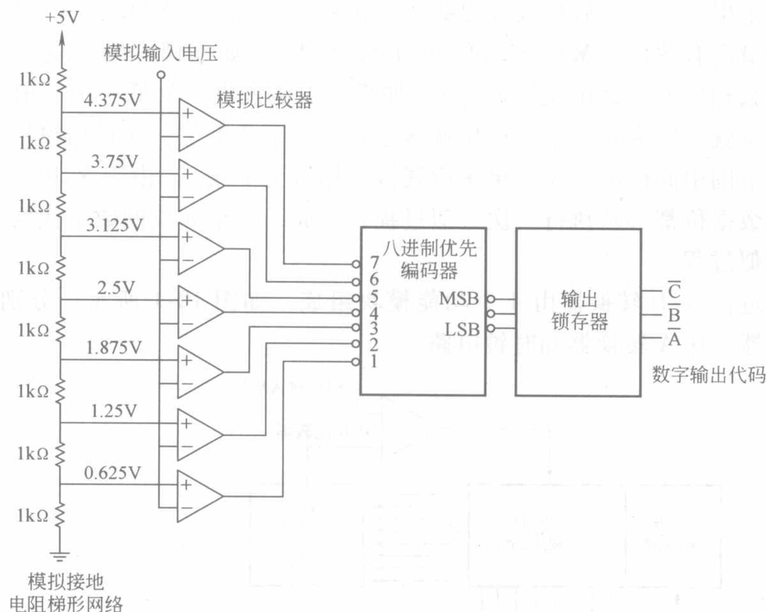


图 10-2 快速 A/D 转换器功能模块（来源：Garrod，S. and Borns，R. 1991. *Digital Logic: Analysis, Application, and Design*，p. 928. 该著作由 Saunders 大学出版，再版已经过允许）

一个 8 位的快速 A/D 转换器需要 255 个比较器。高分辨率 A/D 比较器的成本随电路复杂度的上升而增加，也随模拟转换器数量的增加而增加 2^n-1 倍。为了采用低成本的转换器，部分制造商推出了改进型的快速 A/D 转换器，该转换器的 A/D 转换过程分为两步，以此来减少所需电路的数量。这种改进型的快速 A/D 转换器也称为“半-快速 A/D 转换器”，因为它同时只执行一半的转换过程。

5. A/D 转换器 IC

A/D 转换器 IC 可以分为一般用途、高速、快速以及采样 A/D 转换电路。一般用途转换电路通常具有低速和低成本的特点，其转换时间范围为 2μs～33ms。

这些器件中常用的A/D转换技术包括逐次近似、跟踪和积分。一般用途的A/D转换电路通常具有连接到简化微处理器接口电路的控制信号。这些IC非常适合过程控制、工业以及设备应用，也可用于环境监控，如地震监控、海洋环境监控以及污染监控。

高速A/D转换电路的转换时间范围为400ns~3μs，其高速性能可以通过逐次近似技术、改进型Flash技术和统计分支A/D转换技术来实现。高速A/D IC的应用包括快速傅里叶变换（Fast Fourier Transform，FFT）分析、雷达数字化、医学设备以及多路复用数据采集器。有些A/D IC具有非常高的线性特征，可应用于特定用途，如数字频谱分析、振动分析、地质研究、声纳数字化以及医学成像。

快速A/D转换器的转换时间范围为10~50ns。快速A/D转换技术使得这些IC应用于很多专用的高速数据采集电路中，如TV视频数字化（编码）、雷达分析、瞬间分析、高速数字示波器、医学超声波成像、高能物理以及机器人视觉成像应用。

采样A/D转换器的IC中有一个采样-保持放大器，这就表示不再需要外部的采样-保持电路了。这种A/D转换器的输出信号频率范围为35kHz~100MHz。系统的速度取决于采样A/D转换器的A/D转换技术。

A/D转换器的IC通常可以产生并行的或串行的数字代码格式，有些IC还可以同时提供这两种格式。数字输出必须兼容标准的逻辑系列，以便与其他数字系统接口进行连接。另外，有些A/D转换器IC具有内置的模拟多路复用器，因此这些转换器可以接收多路模拟信号。

PCM IC是高精度的A/D转换器。PCM IC通常称为“PCM编译码器”，因为它同时具有编码和译码的功能。编译码器的编码部分执行A/D转换功能，而译码部分则执行D/A转换功能。数字代码的格式通常是串行数据流，这样便于与数字传输和多路复用系统的接口进行连接。

PCM编码过程中，模拟信号首先被采样、量化，然后编码成数字代码。PCM IC中包含了逐次近似技术或其他技术，用来实现PCM编码。另外，PCM编译码器还可以采用非线性数字压缩技术，例如，如果要将输出数字代码的位数降到最低，可以采用压缩扩展。压缩扩展是一种对数技术，用来将代码的位数进行压缩；然后使用反向对数功能来将代码扩展到原始位数，并将其转换成模拟信号。压缩扩展主要应用于通信传输系统中，用来降低数据传输速率，而不会降低低振幅信号的分辨率。压缩扩展有两种标准压缩律：A律和μ律。A律压缩通常被欧洲使用，而μ律则主要在美国和日本使用。线性PCM转换通常用于高保真度的音频系统中，用来保护音频信号在整个模拟范围内的完整性。

数字信号处理（Digital Signal Processing，DSP）技术提供了另一种类型的A/

D 转换器 IC。专用的 A/D 转换如自适应差分脉冲编码调制（Adaptive Differential Pulse Code Modulation，ADPCM）技术、σ 增量调制、语音次波段编码、自适应预示语音编码以及语音辨识等，都可以通过 DSP 系统来实现。有些 DSP 系统需要模拟前置电路，模拟前置电路采用了传统的 PCM 编译码 IC 或 DSP 接口 IC。这些 IC 可以与数字信号处理器连接，构成一个先进的 A/D 应用系统。有些制造商在单芯片 A/D IC 中集成了 DSP 技术，如 Motorola 公司生产的 DSP56ACD16σ 增量调制 IC。

积分 A/D 转换器通常用于那些转换时间很长的情况，例如电压表或传感器（如热电偶）。积分 A/D 转换器产生一个数字代码，该数字代码代表了信号整个时间内的平均值。通过信号平均或积分可以减小干扰。双斜率积分可以通过一个计数器来实现，当输入电压在指定时间间隔 T 内对电容充完电，该计数器计数值就会增加。这是与另一个计数序列进行比较的结果，当参考电压在指定时间（Δt）内对同一个电容放完电时，该计数序列就会增加。充电计数值与放电计数值的比值和输入电压与参考电压的比值成正比。因此，积分转换器可以提供一个数字代码，该数字代码是指定时间内平均输入电压的度量方法。转换精度跟电容以及时钟频率无关，因为它们同时影响着充电和放电的操作过程。充电时间 T 被选作基频时间，该时间将会被丢弃，最大转换速率稍小于每秒 $1/(2T)$ 次转换。虽然这个速率限制了高速数据采集系统中的转换速度，但是对于持续时间较长的低速率变化输入信号应用来说已经足够了。

6. D/A 和 A/D IC 中的接地电路和支路

D/A 和 A/D IC 中需要正确的接地电路和电容性的旁路，以便根据性能规范来工作。数字信号会严重削弱模拟信号；为了防止数字信号产生的电磁干扰，模拟和数字接地电路必须分开，而且在电路板上只能有一个公共点。如果有可能，这个公共点最好连接到供电电源。

当电源连接到 IC 芯片时，或者输入参考信号时，或者为了在模拟输入端减小由数字信号产生的噪声干扰时，就需要旁路电容。每个制造商都会在数据手册中指定推荐旁路电容的位置和数值。在使用时，最好参照制造商的建议，以确保相应的性能。

7. D/A 和 A/D 转换器 IC 的选择标准

目前，有成百上千的 D/A 和 A/D 转换器 IC 可供我们选择，其价格也从几美元到几百美元不等。为了选择合适的转换器，必须基于系统的应用要求、性能要求和成本。为了方便选择合适的转换器，下面列出了必须考虑的几个要点：

1）系统的输入和输出有什么要求？确定所有信号的电流和电压范围、逻辑电平、输入和输出阻抗、数字代码、数据传输速率以及数据格式。

2）系统要求的精确度是多少？确定所有模拟电压范围、动态响应、线性度

以及编码位数的分辨率。

3）系统要求的速度是多少？确定 A/D 系统中模拟输入的最大采样频率、每个模拟信号的编码位数以及 D/A 系统中输入数字代码的变化率。

4）系统的运行环境是怎样的？收集关于温度范围和供电的信息，以便选择转换器，使之准确覆盖系统的工作范围。

D/A 和 A/D 转换器 IC 的最后选择最好咨询厂商以获取他们产品的技术规范。

名词解释

压缩扩展：为了减小信号传输速率，通过在传输前进行压缩，而在接收端进行扩展的过程。它实际上是一种“数据压缩”技术，该技术需要很少的处理过程。

δ 增量调制：一个 A/D 转换过程，其中数字输出代码代表了模拟输入信号的变化或斜率，而不是模拟信号的绝对值。例如，A1 表示输入信号的上升斜率；A0 表示输入信号的下降斜率。采样率取决于信号的导数，因为为了确保性能的可接受性，快速变化的信号必须有一个快速的采样率。

固定参考电压 D/A 转换器：模拟输出与固定（不变化）参考信号电压成正比。

快速 A/D：目前最快的 A/D 转换过程，也称为“并行 A/D 转换”。模拟信号同时被 2^n-1 个比较器评估，以在每个阶段产生一个 n 位的数字代码。由于需要大量的比较器，因此，快速 A/D 转换器的成本就很昂贵。这种技术通常用于数字视频系统中。

积分 A/D：模拟输入信号根据时间被积分，用来产生一个数字信号，该数字信号代表了特性曲线以下的区域或部分。

乘法 D/A 转换器：一个 D/A 转换过程，其输出信号是数字代码与一个模拟输入参考信号的乘积。这就使得模拟参考信号可以由数字代码来度量了。

Nyquist A/D 转换器：模拟信号的最高频率小于 Nyquist 频率，而 Nyquist 频率的值定义为采样频率的一半。如果一个信号的频率高于 Nyquist 频率，那么就会产生失真，称为“混淆失真”。为了防止混淆失真的发生，就需要一个具有平坦通频带和陡峭边缘带的抗混淆失真滤波器。

超采样转换器：采样频率远高于 Nyquist 频率的 A/D 转换器。典型的采样率是 Nyquist 转换器所需采样率的 32 ~ 64 倍。

脉冲编码调制（PCM）：需要 3 步完成的 A/D 转换过程，这 3 步分别为：对模拟信号的采样、量化以及编码成固定长度的数字代码。这种技术通常应用于数

字语音和音频系统中。相反的过程就可以从 PCM 代码中重构模拟信号。这个过程非常类似于其他 A/D 技术，但是应用于语音和音频系统的特定 PCM 电路可以进行优化。

σ 增量调制 A/D 转换：一个超采样 A/D 转换过程，其中模拟信号的采样率远高于（典型的为 64 倍）Nyquist 转换器所需的采样率。σ 增量调制在进行之前要对模拟信号进行求积分。然后对模拟信号的积分进行编码，而不是对模拟信号的变化值进行编码，这就是传统的 δ 增量调制过程。数字采样率精简滤波器（也称为“数字抽选滤波器”）通常用来提供一个输出采样率是 Nyquist 频率两倍的信号。超采样和数字采样率精简带来的结果是相对于 Nyquist 转换器提高了分辨率，减小了失真。

逐次近似：一个 A/D 转换过程，该过程通过 n 个步骤系统地估计了模拟信号，并产生一个 n 位的数字代码。模拟信号被逐次进行比较来产生数字代码，数字代码的第一位是最高有效位。

参考文献

[1] Analog Devices. 1989. *Analog Devices Data Conversion Products Data Book*. Analog Devices, Inc., Norwood, MA.

[2] Burr-Brown. 1989. *Burr-Brown Integrated Circuits Data Book*. Burr-Brown, Tucson, AZ.

[3] DATEL. 1988. DATEL *Data Conversion Caralog*. DATEL, Inc., Mansfield, MA.

[4] Drachler, W. and Bill, M. 1995. New High-Speed, Low-Power Data-Acquisition ICs. *Analog Dialogue* 29 (2): 3-6. Analog Devices, Inc., Norwood, MA.

[5] Garrod, S. and Borns, R. 1991. *Digital Logic: Analysis, Application and Design*, Chap. 16. Saunders College Publishing, Philadelphia, PA.

[6] Jacob, J. M. 1989. *Industrial Control Electronics*, Chap. 6. Prentice-Hall, Englewood Cliffs, NJ.

[7] Keiser, B. and Strange, E. 1995. *Digital Telephony and Network Integration*, 2nd ed. Van Nostrand Reinhold, New York.

[8] Motorola. 1989. *Motorola Telecommunications Data Book*. Motorola, Inc., Phoenix, AZ.

[9] National Semiconductor. 1989. *National Semiconductor Data Acquisition Linear Devices Data Book*. National Semiconductor Corp., Santa Clara, CA.

[10] Park, S. 1990. *Principles of Sigma-Delta Modulation for Analog-to-Digital Converters*. Motorola, Inc., Phoenix, AZ.

[11] Texas Instruments. 1986. *Texas Instruments Digital Signal Processing Applications with the TMS320 Family*. Texas Instruments, Dallas, TX.

[12] Texas Instruments. 1986. *Texas Instruments Linear Circuits Data Acquisition and Conversion Data Book*. Texas Instruments, Dallas, TX.

备注

Analog Devices 公司编辑出版了很多技术手册来帮助设计工程师满足他们数据采集系统的要求。在这

些参考资料中，还可以获取更多更深的技术信息。这些出版资料包括：《*Analog-Digital Conversion Handbook*》，该手册由 Analog Devices 公司的工程小组编写，由 Prentice-Hall，Englewood Cliffs，NJ 于 1986 年出版。另外，Analog Devices 公司（Norwood，MA）还出版了《*Nonlinear Circuits Handbook*》、《*Transducer Interfacing Handbook*》和《*Synchro and Resolver Conversion*》等刊物和手册。

工程行业杂志和设计出版物上经常有描述最新 A/D 和 D/A 转换电路及其应用的文章。这些出版物包括：《*EDN Magazine*》，《*EE Times*》和《*IEEE Spectrum*》。而与研究相关的主题，读者可以参考“*IEEE Transactions on Circuits and Systems*”和“*IEEE Transactions on Instrumentation and Measurement*”。

第 11 章　专用集成电路

Constantine N. Anagnostopoulos
Paul P. K. Lee

11.1　引言

专用集成电路（Application Specific Integrated Circuit，ASIC）又称为“定制集成电路”，是指通过在单块芯片或少量芯片上集成大量功能性电路来实现以下目标的芯片：实现标准元器件无法实现的功能；提高电路的性能；降低体积、重量以及功率要求，并提高给定系统的可靠性。

ASIC 可以分为以下 3 种：全定制、半定制、可编程逻辑器件（PLD）。

在设计定制 IC 芯片的过程中，第一步要做的就是定义芯片所要实现的功能，这个过程是在系统规划时完成的。在系统规划的过程中，系统工程师和 IC 设计师会制定一些原始决策。根据这些决策，我们可以通过使用标准、通用的元器件以及定制 IC 来实现电路的各种功能。在经过反复讨论之后，每个定制芯片要实现的功能就被确定下来了。

通常，一个给定的功能都可以通过各种不同的方式来实现。例如，许多功能既可以在模拟域实现，也可以在数字域实现。如果在数字域里实现，那么我们可以选择不同的执行策略。例如，一个延迟功能既可以通过一个移位寄存器来实现，也可以通过随机存取存储器来实现。因此，在设计定制 IC 芯片过程的第二步中，系统工程师和 IC 设计师必须确定芯片所要实现的各个功能的执行方式。给定芯片上一个功能的实现方式就决定了该芯片的性能描述。

第三步，选定每个定制芯片的设计方法。根据芯片的实现方式，可以选择全定制、半定制或用户可编程逻辑器件等方法来完成芯片的设计。

设计方法的选择有一个很重要的注意事项，即成本和设计周期（Fey 和 Paraskevopoulos，1985）。例如，一个全定制的设计方法占用的设计时间最长，而且每块芯片的成本也非常高，除非芯片的产量也很大，达到每年数十万块芯片，否则不采用。设计时间最短的方法是采用可编程逻辑器件，但是每块芯片的成本是最高的。对于标准的和小批量的系统来说，采用可编程逻辑器件是最好的选择。

11.2　全定制 ASIC

在一个典型的全定制 IC 设计中，芯片上的每个元器件都是为专用的芯片设计的。当然，根据常识，如果之前设计的元器件工作正常的话，那么我们会尽可能继续使用这些元器件。

全定制的设计方法可以减小芯片尺寸，也可以实现半定制或标准 IC 无法实现或无法优化的功能。减小芯片尺寸可以提高芯片的生产效率以及每个晶片上芯片的数量。以上这两个特点都可以降低每个芯片的成本。

生产效率由下面的公式定义：

$$Y=[(1-e^{-AD})AD]^2$$

式中，A 是指芯片或冲模面积；D 是指每个晶片在每平方厘米上的平均缺陷数量。

每个晶片上冲模的数量为

$$N=[\pi(R-A^{1/2})^2]/A$$

式中，R 是指晶片的半径；A 是指冲模的面积。

如果当冲模还在晶片上时就对其进行测试，并对性能良好的冲模进行封装后再重新进行测试，那样就会增加额外的成本。性能良好的冲模必须进行可靠性测试，以获取正常操作条件下芯片的期望寿命（Hu，1992）。

全定制 ASIC 的制造是在硅铸造厂里完成的。其中有些铸造厂商是不对外的，也就是说，他们只为自己公司的系统部门制造器件而其他铸造厂商也只是对外开放一部分的生产线。但也有专门为外面客户服务的铸造厂商，这些铸造厂商通常还提供设计服务。但是，如果用户只对自己的设计感兴趣，那么铸造厂商可以为用户提供一些关于每个设计过程中的有效设计规则。这些设计规则详细地描述了各种制造工艺，包括 CMOS、双极性器件、BiCMOS 以及 GaAs，并指定了最小尺寸；该尺寸可以在晶片上定义各层的参数、各个动态器件的 SPICE 参数（Vladimirescu 等，1981）、无源器件的参数值范围以及其他规则和限制。

设计全定制芯片是一项艰巨的任务，只能由专门的 IC 设计师来完成。通常，设计全定制芯片需要一个设计小组来共同完成，以节省设计时间（Fey 和 Paraskevopoulos，1986），因为每个人不可能拥有全部的设计经验。同样，还需要功能强大的计算机硬件和软件。通常来说，计算机辅助设计（CAD）工具的功能越强，设计出来的芯片初次工作成功的可能性就越高。如果一个全定制芯片的设计时间和制造周期都很长，而且成本也很高，那么我们必须注意（同时也是很重要的一点）尽可能使用自动设计，并尽可能保证设计规则检测的有效性，以及电路仿真的完整性和精确性。

在设计阶段，我们必须在芯片附加电路的集成上下很大的功夫，以便在芯片生产出来后用来辅助检验芯片工作是否正常，或者用来确定导致芯片无法工作的

原因（元器件或小部分电路），或者确定生产过程中出现的错误。对于数字全定制电路来说，大量的易测性技术已经被开发出来了（Williams 和 Mercer，1993），大部分的这些技术都是自动的，而且在给定合适的软件时，很容易集成到设计中。而对于模拟电路来说，还没有可接受的、通用的易测性方法，对于每个电路专用的测试方法还有待开发。

一个特殊的全定制 ASIC 设计技术的选择取决于该芯片所要实现的功能、该芯片的性能规范以及理想的成本。

11.3 半定制 ASIC

半定制 ASIC 和全定制 ASIC 的主要区别是半定制 ASIC 的基本电路功能模块（无论是模拟还是数字）事先已经被设计好，而且可以正常工作。这些基本电路位于 CAD 系统的基本电路元器件库里面，用户可以任意从电路元器件库中挑选所需的元器件来构建电路并进行布线。另外，电路的仿真水平也比 SPICE 高很多，因此，设计师就不需要对半导体或元器件的物理结构非常熟悉。

半导体 IC 的设计是通过门阵列、标准单元、模拟阵列、功能模块以及 PLD（如 FPGA）来实现的。我们注意到，这些产品在工业中并没有给出一个很标准的命名规则。通常，不同的制造商使用不同的名称来描述基本上类似的产品。门阵列、标准单元和 FPGA 通常用于数字电路设计；而模拟阵列通常用于模拟电路设计；功能模块在两种电路中均可使用。

1. 门阵列

门阵列由规则排列的晶体管组成，通常是由两对 N-沟道晶体管和 P-沟道晶体管排列而成，这是形成 NAND 门电路的晶体管最低数量要求，也是焊接点的固定数量要求；同时，每个晶体管还搭配一个输入输出（I/O）缓冲器。门阵列之间的主要区别特征是它们其中一部分是制造商预先制造好的，设计师的定制工作只需要进行最后焊接和金属层的设计。预先制造好部分元器件可以缩短交货的时间并减小成本，尤其是对那些标准器件来说，这一点尤为突出。

图 11-1 给出了具有 2048 个门电路的 CMOS 门阵列示意图。该器件由 16 列晶体管组成，每一列晶体管包含 128 对 N-沟道晶体管和 P-沟道晶体管。在各列之间，有 18 个垂直的导线沟道，每个沟道包含 21 条导轨。沟道中没有有源元器件。该元器件中的每一对门电路包含 4 条水平导线轨道，因此，整个阵列包含 512 条水平轨道或通道。根据经验，在给定门电路数量的门阵列中，每个门电路包含的导线轨道或通道数 R 可以由下面的公式来计算（Fier 和 Heikkila，1982）：

$$R = 3CG^{0.124}$$

式中，C 是指每个门电路的平均连接数量；G 是指阵列中的门电路数量。

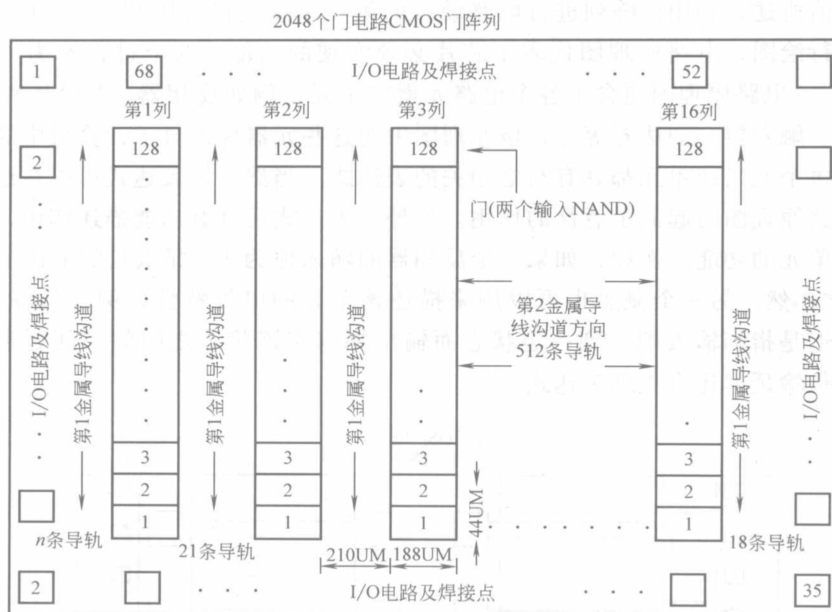


图 11-1　具有 2048 个门电路的 CMOS 门阵列示意图

对于一个双输入的 NAND 门电路来说，如图 11-2 所示，连接数量 C 为 3，两个输入端分别为 A 和 B，输出端为 $\overline{AB}$；对于 2048 个门电路的门阵列来说，根据前面的公式可以计算出 $R \approx 23$。那么，在该器件中，每个门电路就可以提供 23 条导线通道。

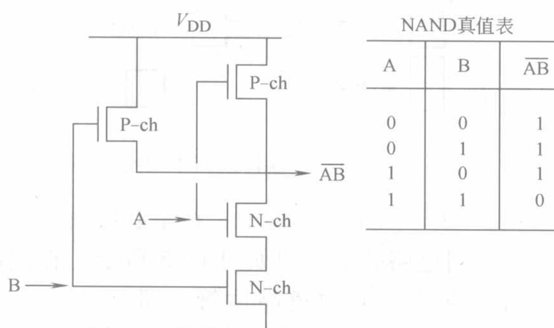


图 11-2　CMOS NAND 逻辑门电路电气示意图及真值表

在该元器件的周边排列了 68 个焊接点。其中，8 个点必须用于接电和接地。根据 Rent 规则，内部门电路进行有效通信所需的 I/O 点数量可以由以下公式来计算：

$$P = CG^a$$

式中，P 是指输入输出点的总和；C 是指每个门电路的平均连接数量；G 是指阵列中的门电路数量；a 是 Rent 指数，其值范围为 0.5 ~ 0.7。

对于大规模集成（LSI）电路来说，$a = 0.46$。假设 a 为该值，加上 $G = 2048$，$C = 3$，那么 P 就等于 100。因此，该门阵列就只拥有 60 个 I/O 焊接点，这样在很多设计中会受到焊接点数量的限制。

如前所述，利用门阵列进行电路设计的第一步就是设计师利用 CAD 系统的辅助进行绘图，电路原理图代表了芯片必须实现的功能，这个过程称为“原理图获取”。电路原理图包含了各个电路元素或单元，例如反相器、NAND 和 NOR 门电路、触发器、加法器等等，该原理图中的这些元器件来自于计算机中的元器件库。每个元素或单元都具有与之相关的表达式。当然，该表达式可以在绘制完整的电路原理图时起到示意性的作用。另外，表达式还具有功能描述作用，可以说明该单元的功能。例如，如果一个反相器的输入值为 1，那么它的输出值就是 0，反之亦然。另一个概念也可以用来描述该单元的电气特性，例如传输延迟，传输延迟是指当输入端达到一定状态而输出端响应该状态之间的时间延迟。另外，该传输延迟还有物理表达式。

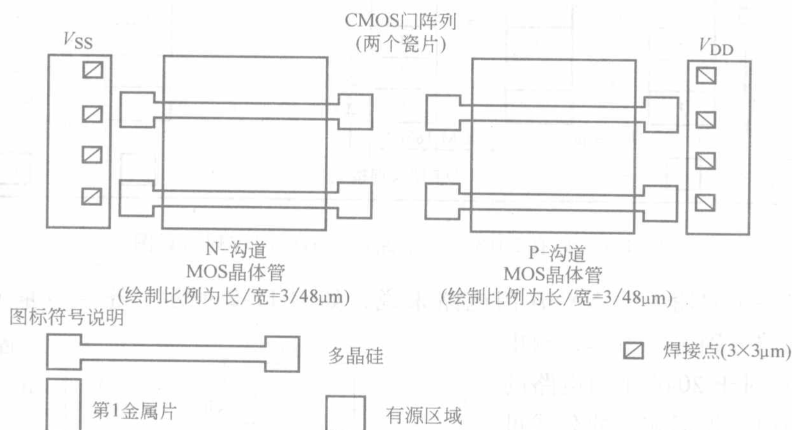


图 11-3 器件纵列中 2048 个未定性门电路之一的示意图

如图 11-1 所示的阵列中，在定制层被布置到元器件上之前，2048 个门电路中每一个门电路的示意图如图 11-3 所示。在图11-3中，左边有两个 N-沟道晶体管，右边有两个 P-沟道晶体管。同时，在左边还有一个接地或 V_{SS} 总线，在右边还有一个 V_{DD} 电源线。图 11-4 给出了图 11-3 电路结构的等效电气原理图。图 11-5 只给出了一对 N-沟道晶体管和 P-沟道晶体管的示意图以及元器件区域的定位，在该区域可以布置其中一个定制层（焊接点）。元件库中器件单元的物理表达式包含了未定性晶体管如何在元器件纵列中进行连接的信息。

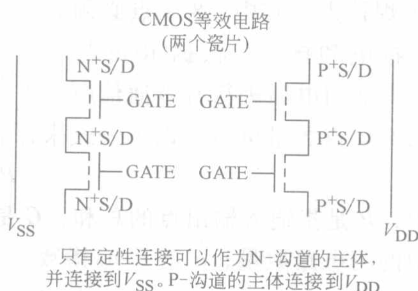


图 11-4 门阵列元器件纵列区域中未定性门电路的等效电气示意图

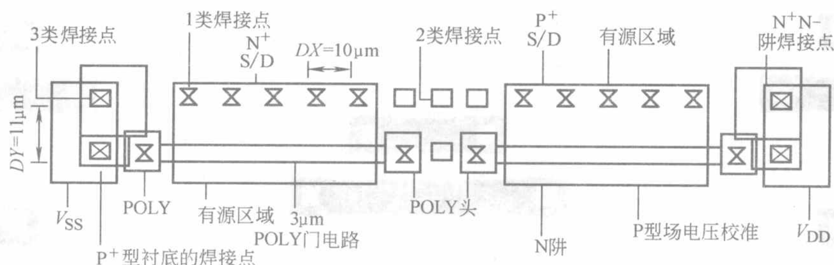


图 11-5　元器件纵列中一对 N-沟道晶体管和 P-沟道晶体管的示意图

注：图中给出了焊接点的分配方案

图 11-6 给出了门阵列中 16 个器件纵列之一的部分示意图。图 11-7 给出了焊接口的位置（即图中的白色方块）以及反相器的金属导片位置（即水平的黑色横条）。而实际的反相器电路是通过将反相器单元布置在各列之间来实现的，如图 11-8 所示。图 11-9 给出了门阵列的电气连接示意图。其中，图 11-9b 所示的是反相器的参考电路原理图。我们可以注意到，在图 11-8 和图 11-9 中，顶部的多晶硅门电路以及底部的 N-沟道晶体管连接到 V_{SS}，对应的 P-沟道晶体管连接到 V_{DD}。这样，就可以将反相器与另一个处于其上面或下面的逻辑门电路进行电气隔离。

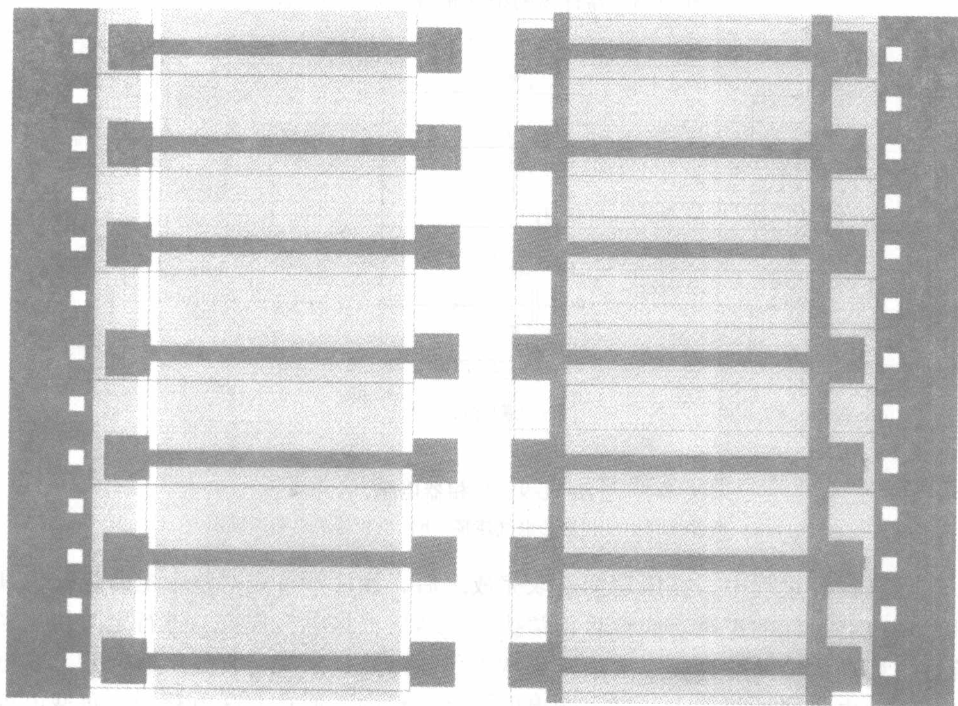


图 11-6　器件纵列的部分示意图

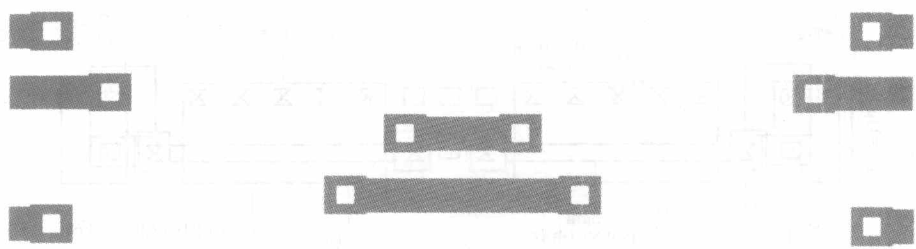


图 11-7　单反相器单元示意图

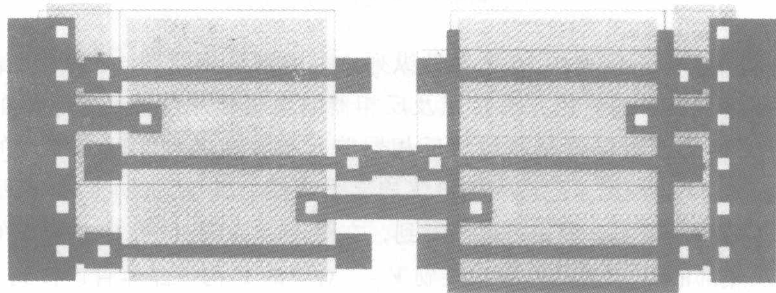


图 11-8　器件纵列中反相器的完整示意图

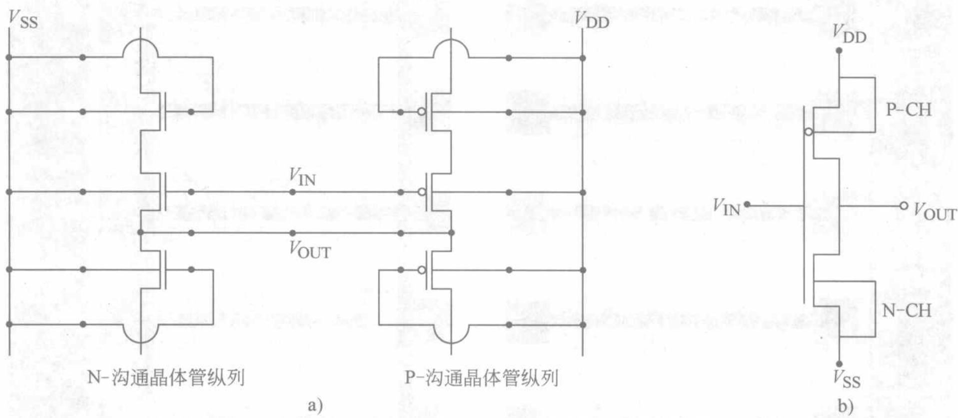


图 11-9　反相器电路

a）器件纵列中反相器的电气连接　b）反相器的电路原理图

回到设计过程中，在原理图获取完成之后，设计师首先对整个电路进行仿真来验证电路设计的逻辑功能能否正常执行；然后根据电气规范，进行定时仿真来确保电路在所需的时钟频率下可以正常工作。但是，定时仿真还不够完整。在每个单元的电气特性规范中，输入输出电容是假设的；而在实际电路中，负载电容则大不相同。因此，在每个节点实际的电容值确定后，要再进行一次定时仿真。

为了进行定时仿真，首先必须从电路原理图中提取出“网络表”。该网络表包含了关于给定单元节点连接到其他单元节点的准确信息，并被提交到定位和布线程序中。定位和布线程序负责将电路中的所有元素或单元布置在门阵列上，类似于图 11-8，并根据网络表对这些单元进行布线。该程序会重复单元的位置定位和布线过程，直至网络表中的所有单元都布置在门阵列上并按要求连接好。

定位和布线程序中布线的部分称为“电路布线器”。顾名思义，布线器用来连接阵列中的各个单元。同时，门阵列的结构（见图 11-1）由多条垂直的导线沟道和一条水平导线沟道构成，以便可以适用于该类型的布线器。图 11-10 给出了定制之前 2048 个门电路的实际门阵列示意图，而图 11-11 给出了定制电路完成之后的实际门阵列示意图。图 11-12 给出了阵列一角的放大示意图。在该图中，水平颜色较深的线是第一层金属导片，而垂直颜色较浅的线是第二层金属导

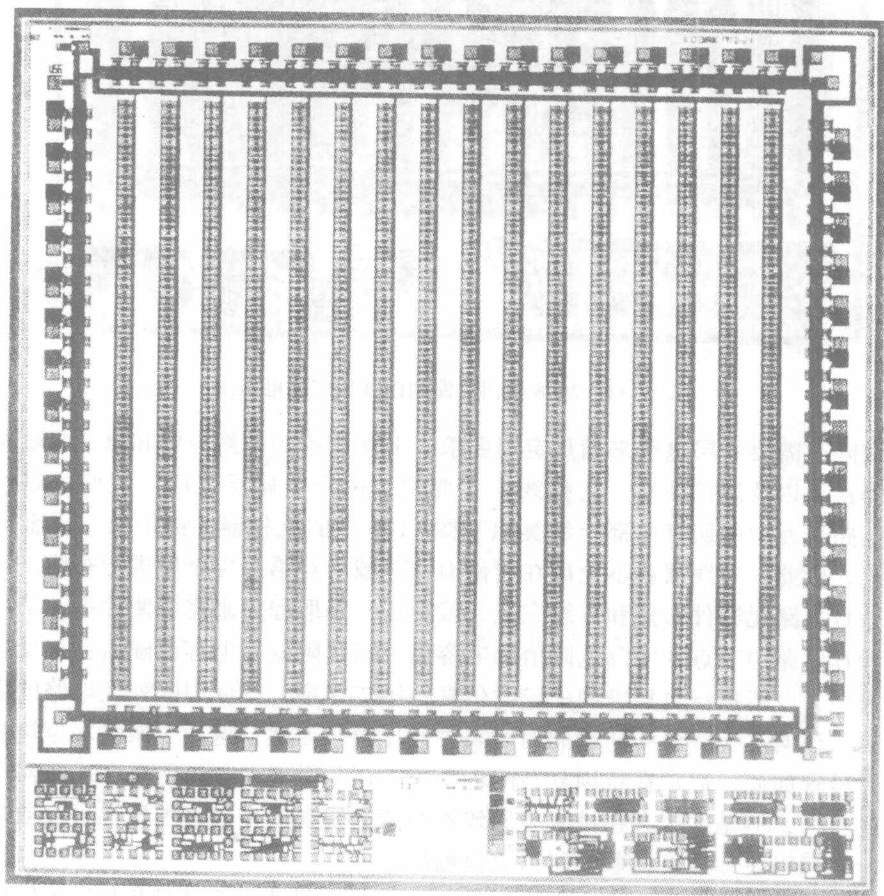


图 11-10　实际 2048 个门电路的门阵列定制前的示意图

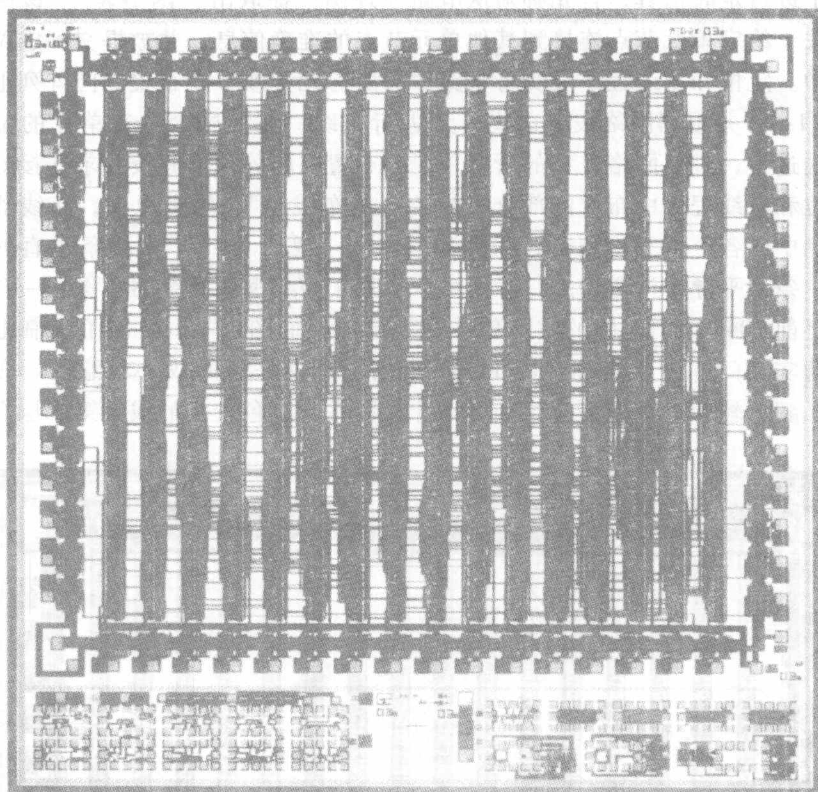


图 11-11 2048 个门电路的门阵列定制电路图

片。该阵列需要 4 层电路来满足定制要求。上面介绍的是第一层和第二层金属导片、焊接点以及连接线路。连接线路主要位于两个金属层之间。我们可以注意到，电路单元中的所有内部导线类似于图 11-7 中的反相器，都由第一层金属导片产生，因此，布线器就不允许在器件面积区域上对第一层金属进行布线。

一旦电路元器件布置和布线完成，就会运行提取程序来完成两件事。第一件事，计算电路中节点的实际电阻值和电容值，该电阻值和电容值随后会被反馈到网络表中，以便进行更加准确的定时仿真；第二件事，提取门阵列上电路的最新网络表，该网络表可以和初始的网络表进行比较。当然，这两个网络表应该是相同的。尽管如此，通常在特别复杂的电路中，还是需要人工控制来完成元器件的布置或布线，但是这也会在排版中导致不经意的错误产生。这种错误可以通过比较原始的网络表和最新提取的网络表检测出来。

电路设计完成之后，就会产生两组测试矢量，即输入矢量组和相应的输出矢量组。每个测试矢量由多个 1 和 0 组成。输入测试矢量的分量数量和输入端引脚

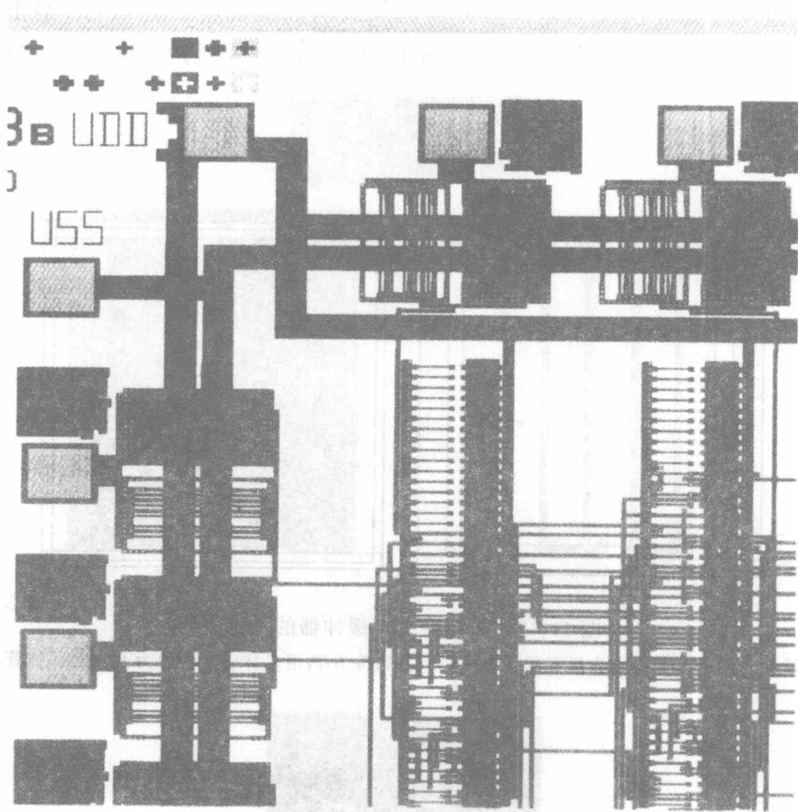


图 11-12　ASIC 一角的示意图

数量相同，类似的，输出测试矢量的分量数量和输出端引脚数量相同。

输入测试矢量不同于实际运行中芯片所看到的。因此，我们的目标就是挑选一组输入矢量，该输入矢量在应用到芯片上时，可以导致内部节点至少改变一次芯片的状态。定时验证程序就是用来完成这项任务的。当每个输入矢量被应用时，该程序将保持固定节点的电路并保留输出矢量。当应用到精选的器件中时，就会顺序应用相同的输入测试矢量，并可以捕捉到相应的输出测试矢量，最后将该结果和定时仿真的结果进行比较。如果这两个结果非常匹配，该器件就是合格的。

以上测试的目的是检测内部门电路运行是否正常（由于在制造过程中可能会产生缺陷）。通常，如果一个测试过程可以导致 90% 的内部节点发生状态转换，那么该器件就可以认为能提供足够的工作范围。而理想的是具有 100% 的工作范围，但是通常这是不实际的，因为完成测试将会占用一定的时间。最后，由于测试是在芯片正常运行的条件下进行的，因此，只执行最少的功能性测试。

如同器件纵列中任何逻辑门电路都可以放置在任意纵列中一样，每个焊接点

都可以通过 4 个定制层成为输入点或输出点或电源接入点。图 11-13 给出了预先制造过程末端的 I/O 缓冲器示意图；图 11-14 给出了焊接点和金属模的示意图，

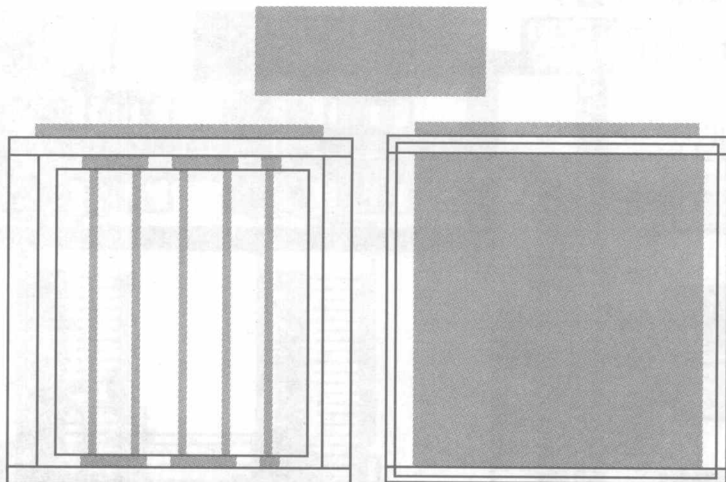


图 11-13　未定型 I/O 缓冲器的示意图

注：在图中，焊接点为顶部的矩形样式，左边为 N-沟道晶体管，右边为 P-沟道晶体管

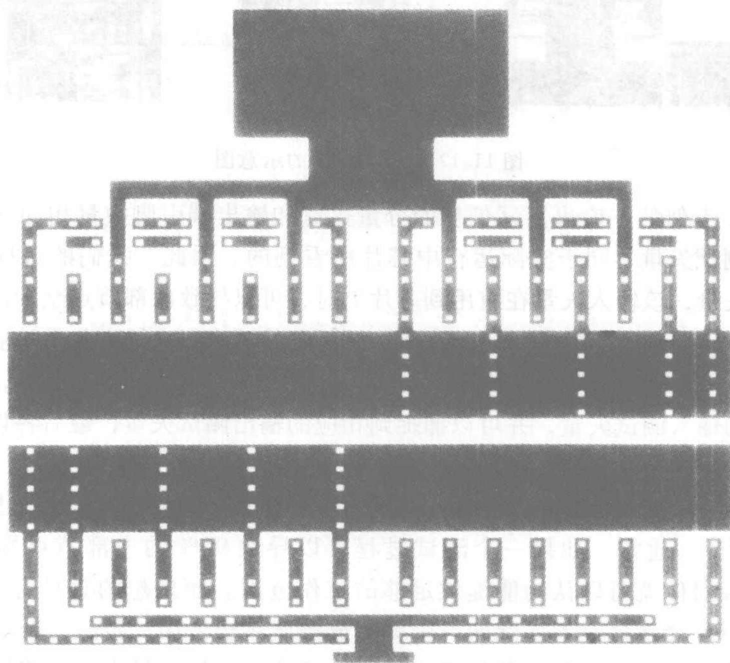


图 11-14　输出缓冲器库单元的接触点和金属模示意图

其中未定型的焊接点作为反相输出缓冲器；图 11-15 给出了以上两者的合并示意图；图 11-16 给出了输出缓冲器的电路示意图和焊接点作为输入缓冲器的示意图；图 11-17 给出了该输入缓冲器的示意图。

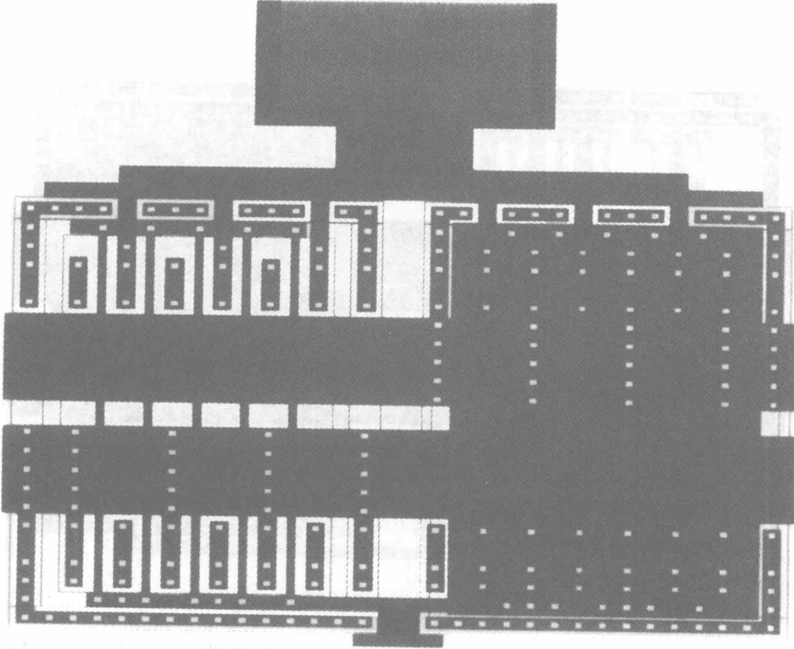


图 11-15　全反相输出缓冲器的完整示意图

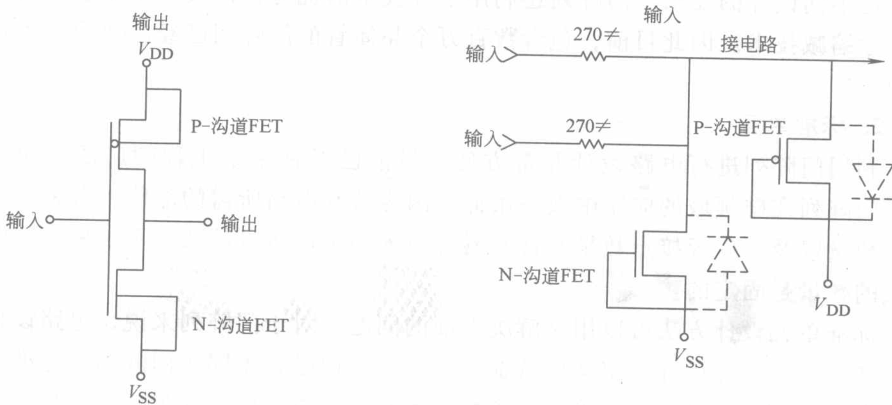


图 11-16　输入和输出缓冲器的电气示意图

注：鉴于与 TTL 的输入或输出端兼容或其他原因，缓冲器只能使用 5 段 N-沟道晶体管或 P-沟道晶体管

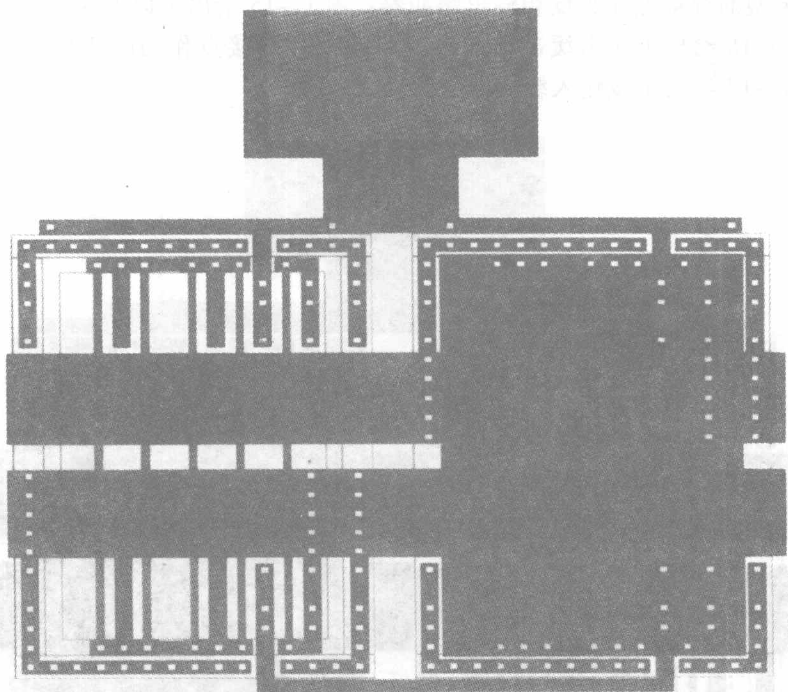


图 11-17 单放大器输入缓冲器的完整示意图

这里介绍的门阵列结构不是惟一的，目前已经开发出了许多不同的结构设计，并得到了广泛应用。对于每种结构来说，需要不同的元器件定位和布线规则来满足不同设计的要求。门阵列还利用了硅技术的优势，硅技术主要包含的是最小尺寸缩减技术。因此目前，包含数百万个晶体管的门阵列已经广泛投入实际应用中了。

2. 标准单元

利用门阵列进行电路设计非常方便，目前已经被系统工程师广泛采用。但是，门阵列在硅领域的应用中效率很低，因为特殊电路所需的晶体管数量、布线路径数量以及 I/O 焊接点数量经常和器件或阵列中焊接点的数量不匹配，其中焊接点的数量是固定的。

标准单元设计方法可以用来解决上面的问题。对于门阵列来说，电路设计师可以使用一个预先设计并定义好特征的单元库。但是，不同于门阵列，标准单元芯片在设计时只需考虑晶体管的数量和由特殊电路指定的 I/O 焊接点，从而使芯片的成本可以更低，尺寸可以更小。但是，标准单元电路的制造必须从加工过程开始。该过程有两个主要缺点：第一，标准单元的制造周期大约为数周，比门阵列的制造周期长；第二，用户必须为整个晶片的制造付费，而门阵列中只有一小

部分晶片需要用户定制。这一点就从根本上增加了元器件的成本。

由于标准单元可以自动定位和布线，因此标准单元的设计必须有一定的规范，例如统一的重量、预先指定的时钟模块和电源线位置、预先定义的输入输出端位置以及其他要求。图 11-18 给出了标准单元库中的 3 个常见标准单元的示意图；图 11-19 给出了一个完整的标准单元芯片，其外形非常类似于门阵列，但它不包含无效的晶体管或 I/O 焊接点，同时该芯片还具有最少的布线通道。

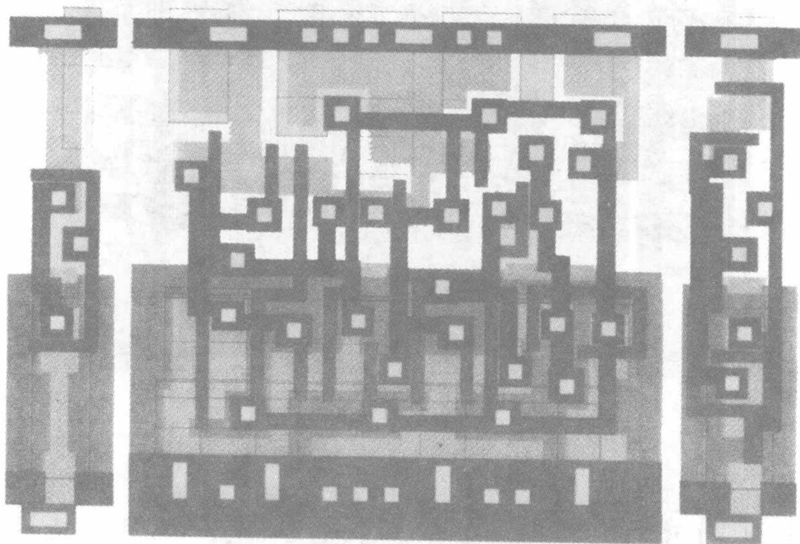


图 11-18　3 个标准单元

注：左边是一个双输入 NAND 门电路，中间是一个 D 型触发器，右边是一个单反相器

3. 功能模块

元器件布置和布线软件的优势使得功能模块设计技术成为了可能。这里，标准单元不再是指简单的逻辑门了，而是指完整的功能模块。该标准单元可以是模拟的也可以是数字的，其外形尺寸可以是任意大小的。图 11-20 给出了一个数字功能模块设计的示例。在该图中，3 个白色的模块是连同几行标准单元自动进行定位和布线的。在图 11-21 中，位于芯片中心的小型线性电荷耦合器件（Charge Coupled Device，CCD）图像传感器与位于传感器下面的其他模拟电路放置在相同的芯片上，该芯片包含一个标准单元模块，用来产生芯片运行所需的所有逻辑功能。在该芯片中，必须注意各个功能模块的定位和布线，因为模拟电路对噪声、温度和其他因素比较敏感。因此，设计师必须认真考虑各个功能模块的定位和布线的输入点，必要时可以利用软件来完成。

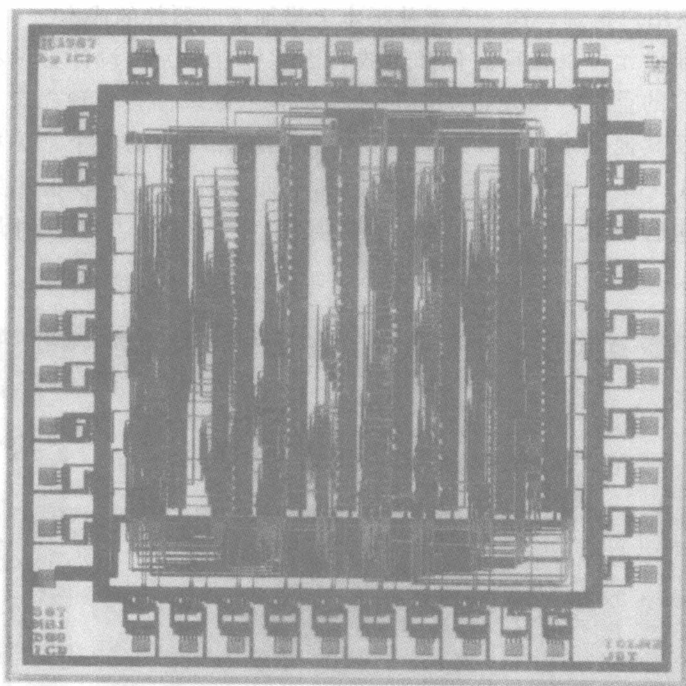


图 11-19 完整标准单元 ASIC 的示意图

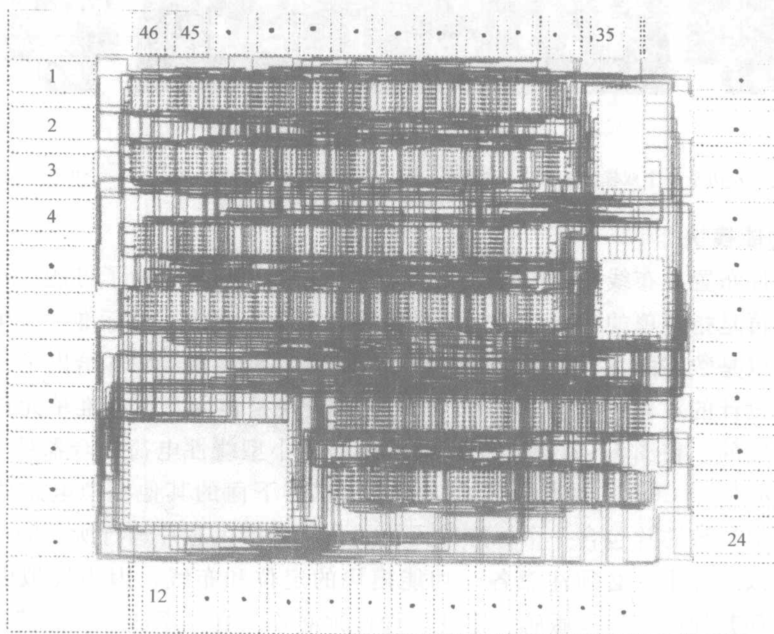


图 11-20 ASIC 的数字功能模块设计示意图

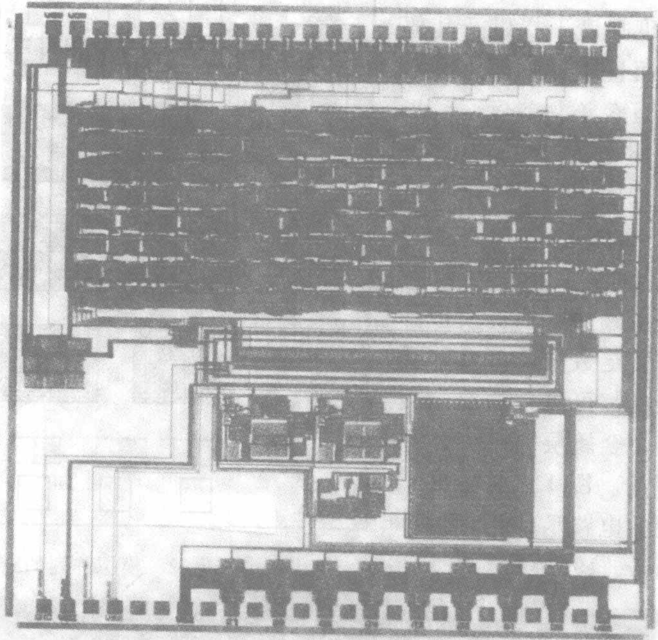


图 11-21　ASIC 的混合（在同一块芯片上集合了模拟电路和数字电路）功能模块设计示意图

通常，数字功能模块是通过一系列的软件来实现的，这些软件称为“硅编译器”或者“合成工具”。通过这些工具，电路设计师可以很轻松地利用特别开发的计算机语言（称为“硬件描述语言（Hardware Description Language, HDL)”）来描述电路的功能，这个过程只是描述电路的性能，而不是构建实际的门电路。这些工具可以模拟任意大小的存储器功能模块、任意长度的加法器或乘法器、PLA 以及许多其他电路。

现在可以很清楚地注意到，半定制设计方法很注重软件工具的使用，这些工具避免了单调乏味的手工排版中的错误倾向，并提供了准确的电路仿真，有助于提高电路的可靠性。其功率和混合器件的使用使得集成电路设计中的复杂度更加明显。

4. 模拟阵列

模拟阵列通常由双极性工艺制造，可用于高性能模拟电路的制造。类似于门阵列，这些模拟阵列可以预先制造成焊接模块，同时可以根据预先设计和特征单元库来设计。但是，由于模拟电路的不可预测性，设计师通常设计自己个性化的单元模块或对已有的单元进行改进。不同于数字电路，模拟电路的排版是人工完

成的，因为模拟电路对温度梯度、电源电压、色度亮度干扰以及其他因素很敏感，而且现有软件还不足以将所有这些因素考虑完全。

模拟阵列的结构类似于“瓷片”的形状，如图 11-22 所示。所有的“瓷片”都是相同的。在每个“瓷片”中包含了很多晶体管、电阻以及各种型号的电容。模拟阵列同时还包含固定数量的 I/O 焊接点，每个焊接点都可以根据定制设计要求，以 4 个金属层为基础作为输入或输出缓冲。图 11-23 给出了模拟阵列实现的电路示意图（Boisvert and Gaboury，1992）。

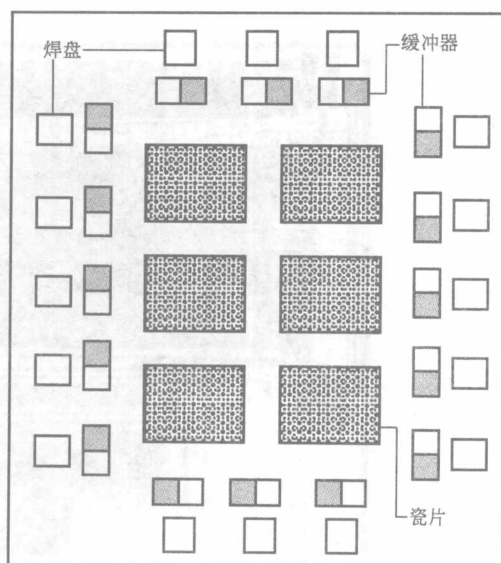


图 11-22 模拟阵列的结构

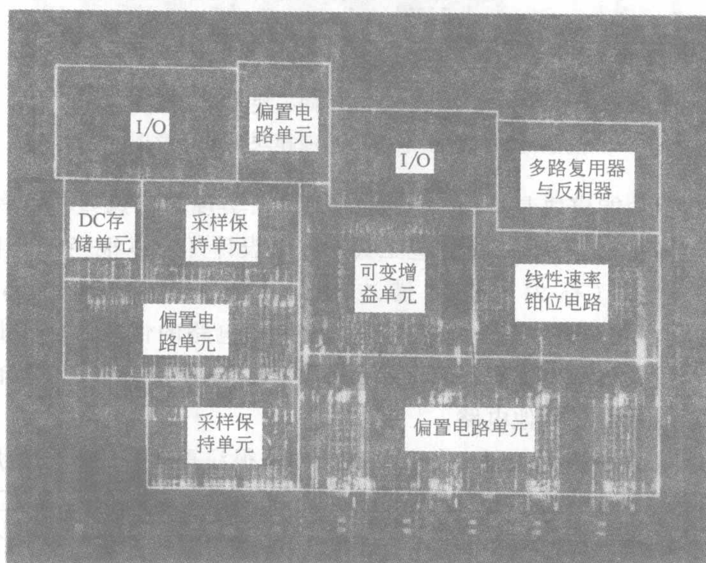


图 11-23 利用模拟阵列实现的电路示意图（Boisvert and Gaboury，1992）

名词解释

ASIC：“专用集成电路”的英文简称，该集成电路的另一个名称是“定制集

成电路"。

功能模块：利用该技术设计的 ASIC 芯片集成度比门阵列或标准单元高，因为功能模块可以比简单逻辑阵列实现更多复杂的功能。

门阵列：该芯片包含了未定型的晶体管阵列，而且可以预先制造成一定的规模，可用于专用电路的制造。

一次性工程（Nonrecurring Engineering，NRE）：可以同时满足制造商利润和 ASIC 用户制造成本的要求。这些成本包括设计成本、掩模制造成本、制造晶片的成本以及封装和测试元器件的成本。

标准单元：用来实现 ASIC 的一种设计方法。相比门阵列，标准单元的硅利用率更高。

原理图获取：通常通过该过程，在计算机的辅助下，利用计算机单元库中的元器件可以以电气原理图的形式获取芯片的功能。

硅编译器、合成工具：可以构建 ASIC 的软件程序。ASIC 的功能不再由电路原理图来描述，而是以一种特定的高级计算机语言来描述，该计算机语言通常称为"硬件描述语言（HDL）"。

测试矢量：由成对的输入端和输出端构成的测试方案。每个输入矢量都是惟一的一组 1 和 0，该 1 和 0 用于芯片的输入；相应的输出矢量是由芯片输出端产生的一组 1 和 0 构成。

参考文献

[1] Boisvert, D. M. and Gaboury, M. J. 1992. An 8-10-bit, 1-40MHz analog signal processor with configurable performance for electronic imaging applications. In *Proceedings IEEE International ASIC Conference and Exhibit*, pp. 396-400. Rochester, NY.

[2] Fey, C. F. and Paraskevopolouos, D. 1985. Selection of cost effective LSI design methodologies. In *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 499-496 . Rochester, NY.

[3] Hu, C. 1992. IC reliability simulation. *IEEE J. of Solid State Circuits* 27 (3): 241-246; see also Proceedings of the Annual International Reliability Physics Symposium.

[4] Ting, G., Guidash, R. M., Lee, P. P. K., and Anagnostopoulos, C. 1994. A low-cost, smart-power BiCMOS driver chip for medium power applications. In *Proceedings ASIC Conference and Exhibit*, pp. 466-469. Rochester, NY.

[5] Vladimirescu, A. et al. 1981. *SPICE Manual.* Dept. of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, Oct.

[6] Williams, T. W. and Mercer, M. R. 1993. Testing digital circuits and design for testability. In *Proceedings IEEE International ASIC Conference and Exhibit* (Tutorial Section), p. 10. Rochester, NY.

备注

ASIC 技术领域自 1980 年以来得到了快速发展，关于 ASIC 最新信息的最佳参考资源主要来自以下两

个报道最新技术的会议，分别为：每年5月举办的IEEE特定集成电路会议（IEEE Custom Integrated Circuits Conference，CICC）和每年9月举办的IEEE国际ASIC会议和展览（IEEE International ASIC Conference and Exhibit）。IEEE频谱杂志发行了研光版的会议资料。除了常规的技术会议外，在这两个会议中还有ASIC提供的教育性会议和展览。

另外，读者还可以参考USC/信息科学学会，4676 Admiralty Way，Marina del Rey，CA 90292-6695，联系电话是（213）822-1511。

最后，自1984年以来CCIC中的某些资料每年都会以特殊的形式出版，如“*IEEE Journal of Solid State Circuits*”。

更多的信息，读者还可参考其他一些会议及其学报，例如设计自动化会议（Design Automation Conference，DAC）和国际固态电路会议（International Solid State Circuits Conference，ISSCC）。

第 12 章　数字滤波器

Jonathon A. Chambers
Sawasd Tantaratana
Bruce W. Bomar

12.1　引言

数字滤波器用来消除离散数据中的噪声和多余的信息，改变采样率以及实现其他功能。尽管离散数据的处理方式有很多种（例如，求出平均值、形成矩形图），但是数字滤波器的目标只有一个，即根据离散输入序列 $x(n)$ 来形成一个离散输出序列 $y(n)$。从某些方面来说，数字滤波器的每个输出样值都是根据输入序列计算得到的——事实上不是根据某一个样值计算得到，而是根据很多个样值或者可能根据所有的输入样值计算得到的。那些根据当前输入样值并加上有限个先前输入样值来计算输出样值的滤波器称为“有限脉冲响应（Finite Impulse Response，FIR）滤波器”；而输出样值全部根据先前的输入样值来计算的滤波器称为“无限脉冲响应（Infinite Impulse Response，IIR）滤波器”。本章将讨论 FIR 和 IIR 滤波器的设计和实现过程，并测试在实现这些滤波器时的有限字长运算效应。

12.2　FIR 滤波器

FIR 滤波器是一个线性离散时间系统，该系统的输出值是指大多数当前输入样值和有限个先前输入样值的加权和。时不变 FIR 滤波器具有有限个存储器，其脉冲响应（也就是对离散输入序列的响应，该输入序列的第一个样值都是一致的，而其他样值都是 0）与滤波器的固定加权系数相匹配。而另一方面，时变 FIR 滤波器的工作采样率可能是变化的，而且其加权系数也会随着滤波器应用环境统计属性的变化而相应变化。

1. 基本原理

最简单的 FIR 滤波器可能就是指根据下面的线性常量系数差分等式描述的移动平均数操作：

$$y(n) = \sum_{k=0}^{M} b_k x(n-k), \quad b_k = \frac{1}{M+1}$$

式中，$y(n)$ 为滤波器在整数采样序号 n 处的输出样值；$x(n)$ 为滤波器在整数采样序号 n 处的输入样值；b_k 为滤波器加权系数，$k=0$，1，…，M；M 为滤波器次序。

在实际应用中，输入和输出离散时间信号是以规则的采样时间间隔（T 秒）进行定义的，分别记为 $x(nT)$ 和 $y(nT)$，该采样时间间隔与采样频率（即每秒钟的采样数）有关，相关公式为 $f_s=1/T$。但是，一般来说，如果假设 T 是统一的就更加方便了，这样有效采样频率就是统一的了，而且 Nyquist 频率（Oppenheim 和 Schafer）只有有效采样频率的一半，也就是说当采样频率为 f_s 时，最大模拟频率不会产生混淆失真了。然后，通过乘法直接对该标准化的频率范围即（0，1/2）进行比例划分来得到其他采样频率。

简单移动平均滤波器的输出值是 $M+1$ 个最新的 $x(n)$ 值的平均值。直观一点说，该值就是对平滑输入样值的响应，但是它的操作更适合通过计算滤波器的频率响应来描述。尽管如此，但是滤波器的 z 域表达式类似于模拟滤波器的 s 域（拉普拉斯）表达式。离散时间信号 $x(n)$ 的 z 域转换公式定义如下：

$$X(z)=\sum_{n=0}^{\infty}x(n)z^{-n}$$

式中，$X(z)$ 是 $x(n)$ 的 z 域形式，z 是复数变量。

延时 $x(n)$（即 $x(n-k)$，其中 k 为正整数）的 z 域形式由 $z^{-k}X(z)$ 表示。这个结果可将简单移动平均滤波器输出样值 $y(n)$ 的 z 域形式与其输入样值相关联起来，公式如下：

$$Y(z)=\sum_{k=0}^{M}b_k z^{-k}X(z),\ b_k=\frac{1}{M+1}$$

这样，z 域形式的转移函数即输入和输出样值比为

$$H(z)=\frac{Y(z)}{X(z)}=\sum_{k=0}^{M}b_k z^{-k},\ b_k=\frac{1}{M+1}$$

注意，该转移函数 $H(z)$ 完全由加权系数 b_k 和 z 的值决定，其中 $k=0$，1，…，M；k 的值和滤波器的离散脉冲响应中的相同。离散脉冲响应的有限长度是指滤波器的瞬间响应只持续 $M+1$ 个样值的时间，之后就达到了稳定状态。滤波器的频域转移函数由 $z=e^{j2\pi f}$决定，其中 $j=\sqrt{-1}$，具体公式如下：

$$H(e^{j2\pi f})=\frac{1}{M+1}\sum_{k=0}^{M}e^{-j2\pi fk}=\frac{1}{M+1}e^{-j\pi fM}\frac{\sin[\pi f(M+1)]}{\sin(\pi f)}$$

简单移动平均滤波器的幅值和相位响应（$M=7$）由 $H(e^{j2\pi f})$ 计算得到，如图 12-1 所示。从图中我们可以很清楚地看到，滤波器如同一个具有线性相位响应的粗糙低通平滑滤波器。幅值和相位响应中的采样频率周期是离散时间系统的一项属性。线性相位响应由 $H(e^{j2\pi f})$ 公式中的 $e^{-j\pi fM}$决定，并对应了滤波器中

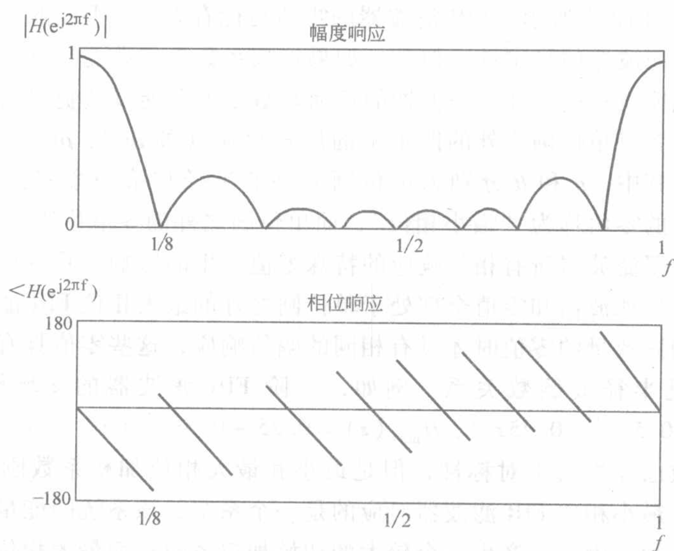


图 12-1　简单移动平均滤波器的幅值和相位响应（$M=7$）

的一个常量 $M/2$ 群时延。幅值属性发生变化时会产生 $\pm 180°$ 的相位中断。加权系数成中心对称的 FIR 滤波器中包含该常量，即频率无关群时延属性，该属性在应用中很重要的，而且在应用时要尽量避免出现时间弥散。例如，在脉冲转换中，要尽量避免产生脉冲变形（Lee and Messerschmit，1994）。

还可以用另一个方式来有效描述 z 域转移函数，如下所示：

$$H(z)=\frac{\sum_{k=0}^{M} b_k z^{M-k}}{z^M}=\frac{b_0 z^M + b_1 z^{M-1} + \cdots + b_{M-1} z + b_M}{z^M}=\frac{N(z)}{D(z)}$$

如上所示，z 域转移函数是两个 M 阶多项式的比值，即 $N(z)$ 和 $D(z)$ 的比值。$N(z)=0$ 时的 z 值称为滤波器的“零值”，而 $D(z)=0$ 对应的 z 值则是滤波器的“极值”。FIR 滤波器的极值是指 z 平面原点处即 $z=0$ 对应的值。零值的位置由加权系数 b_k 决定，其中 $k=0, 1, \cdots, M$。简单移动平均滤波器 z 平面的零值和极值如图 12-2 所示。零值由圆圈标示，和单位圆是重合的，也就是 z 平面上 $|z|=1$ 对应的等值线。该零值准确对应了幅值响应的零值，并由此而得名。相位响应中

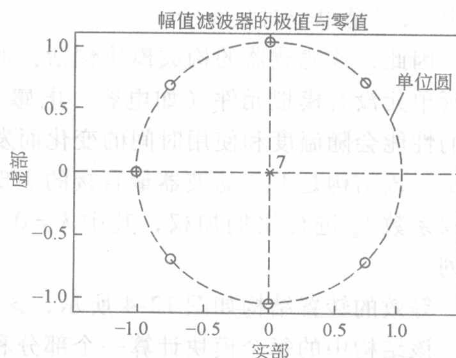


图 12-2　简单移动平均滤波器的零值和极值曲线图（$M=7$）

的中断情况如图12-1所示。FIR滤波器的零值可能存在于z平面的任何地方，因为它们不影响滤波器的稳定性。但是，如果加权系数为实数而且是以中间值$M/2$成对称的，或反对称的，那么滤波器的任何复数零值只能以共轭对的形式出现在单位圆上，或者以单位圆之外的四重根的形式存在（如$\rho e^{j\theta}$、$\rho e^{-j\theta}$、$(1/\rho)e^{j\theta}$、$(1/\rho)e^{-j\theta}$）；其中，ρ和θ分别为单位圆之外的半径和第一个零值的偏转角度数。单位圆上的零值称为“最小相位”，而单位圆之外的零值称为“最大相位”。这种区别描述了滤波器所有相位响应的特殊零值产生的影响。具有单位圆外零值的最小相位FIR滤波器和零值全部处于单位圆之外的最大相位FIR滤波器，只有当它们具有相同数量的零值时才具有相同的幅值响应，这些零值具有相同的偏转角度数，但是半径成倒数关系。例如，二阶FIR滤波器的z域转移函数为$H_{\min}(z)=1+0.5z^{-1}+0.25z^{-2}$，$H_{\max}(z)=0.25+0.5z^{-1}+z^{-2}$。值得注意的是，此时加权系数已经失去了对称性，但是最小和最大相位加权系数刚好成逆序对称。实际上，最小相位FIR滤波器对应的是一个系统，该系统的能量快速从输入端转移到输出端，因此，产生一个较大的初始加权系数；而最大相位FIR滤波器的能量从输入端转移到输出端的速度相对较慢，而且这个较大的初始加权系数也会产生延迟。这类FIR滤波器通常用来模拟多路移动通信传输路径。

FIR滤波器的频率响应特性完全由加权系数b_k的值决定，其中$k=0$，1，…，M，b_k与滤波器的脉冲响应及阶数M相匹配。这些加权系数可以通过多种技术来实现，以满足特定用途的需要。下一节将讨论FIR滤波器实现过程中的有效结构。

2. 结构

FIR滤波器的结构必须实现如下所示的z域转移函数：

$$H(z)=\sum_{k=0}^{M}b_k z^{-k}$$

式中，z^{-k}是单位时延算子。

因此，该滤波器的构成模块包括：加法器、乘法器以及单位时延单元。这些时延单元没有模拟元件（如电容、电感、运算放大器和电阻）的缺陷，模拟元件的性能会随温度和使用时间的变化而发生变化。直接或分支时延形式如图12-3所示，该结构是FIR滤波器最直接的实现方式。输入序列$x(n)$具有时延，并由加权系数b_k进行比例加权，其中$k=0$，1，…，M，最后进行累加以产生输出序列。

等效的转置结构如图12-4所示，该结构更加标准化，更适合集成电路的实现。该结构中的每个模块计算一个部分和，这样在输出端就只需要进行一个简单的加法运算。

其他结构也可以在滤波器的加权系数中实现对称或冗余。例如，具有线性相

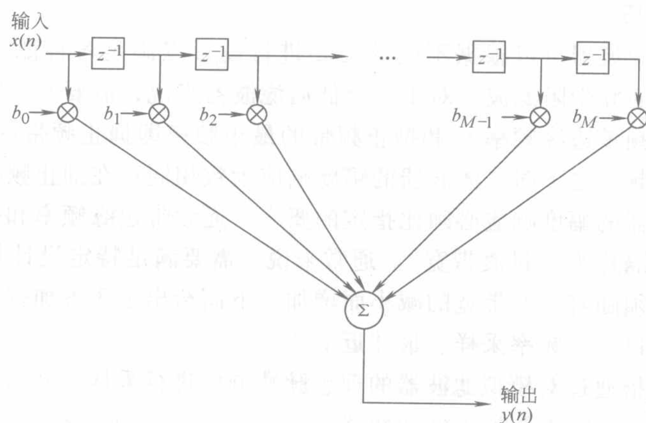


图 12-3 FIR 滤波器的直接实现方式

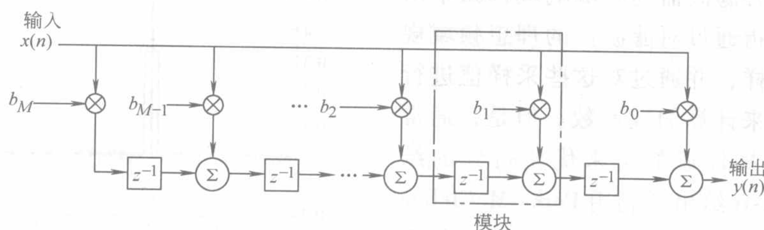


图 12-4 FIR 滤波器的模块化结构

位的 FIR 滤波器就是以其中心加权系数 $M/2$ 为对称的，奇数长度滤波器中的加权系数（M 是偶数）有一半是零，也就是说，理想的单边带 FIR 滤波器只允许 $0<f\leqslant 1/4$ 段的频率通过，并过滤掉 $1/4<f\leqslant 1/2$ 段的频率。因此，通过图 12-5 中所示的格状结构可以实现 FIR 滤波器。格子中的乘法系数 k_j（$j=1$，…，M）在其他 FIR 滤波器中由于加权系数的不同而不尽相同，但是可以通过迭代过程得到。格状结构具有吸引力的地方是可以直接测试其零点是否在单位圆内（即最小相位属性），而且格状结构对量化误差不敏感。这些特性使得格状结构在语音编码中得到了广泛应用（Rabiner and Schafer，1978）。

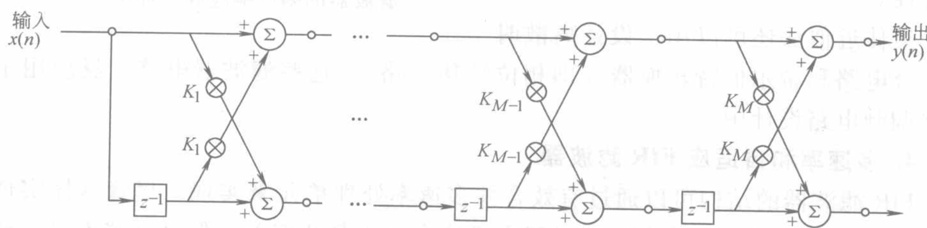


图 12-5 FIR 滤波器的格状结构

3. 设计技巧

线性相位滤波器可以根据不同的需要进行设计以满足各种滤波要求，如低通、高通、带通和带阻滤波。对于一个低通滤波器来说，需要两个频率：通频带最大频率即通频带边缘频率 f_p 和抑止频带的最小频率即抑止频带的边缘频率 f_s。在通频带最大频率之下时，滤波器的幅度响应大致相同；在抑止频带的最小频率之上时，滤波器的幅度响应必须比指定的要小。通频带边缘频率和抑止频带边缘频率之间的间隔称为“过渡带宽”。通常来说，需要满足特定设计规范的 FIR 滤波器的长度必须随着过渡带宽的减小而增加。下面给出了 3 种加权系数设计时的有效技巧：窗口法、频率采样、最佳近似法。

窗口法是指通过对模拟滤波器的理想脉冲响应进行采样，并通过一个平滑的窗口将这些采样值相乘来计算加权系数，以改善滤波器的频域响应；频率采样技术是指通过对滤波器的理想频域规范进行采样，并通过对这些采样值进行反相转换来计算加权系数。但是，通常可以通过最佳近似法来优化得到的结果。图 12-6 给出了利用 Parks-McClellan 算法实现的第 40 个理想单边带 FIR 低通滤波器的示例，同时还给出了理想频域设计的规范。在图 12-6 中，必须注意脉冲响应的零点值。Parks-McClellan 算法使得实际滤波器的幅度响应与理想滤波器的幅度响应的偏差最小。实际滤波器的幅度响应代替了通频带内的理想规范，并超越了抑止频带的理想规范。与理想规范之间的最大偏差在通频带和抑止频带上是均等的，这就是优化方案的特性。

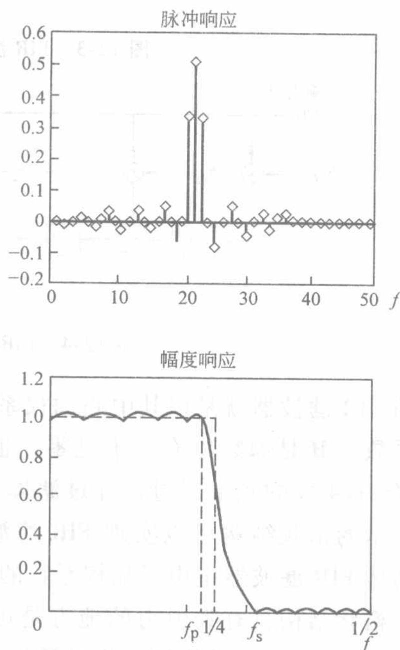


图 12-6 第 40 个理想单边带 FIR 低通滤波器的幅度响应和脉冲响应

最佳近似法还可以用来设计离散时间微分电路和希尔伯特转换器（即相位转移电路），这些滤波器电路广泛应用于数字调制电路设计中。

4. 多速率和自适应 FIR 滤波器

FIR 滤波器的结构可以通过有效合并多速率处理单元来实现。尽管这样实现的滤波器是时变的，而且在处理过程中可能会产生某些偏差，但是这些不足在精细设计时可以避免。假设低通滤波器的 $f_c = 1/8$，那么高于该值的频率就会被过

滤掉。因此，f_c 和 Nyquist 之间的频率范围（即 1/2）就不包含任何有用的信息了。由于避免了混淆失真，因此采样频率可以减小 4 个点。如果在该减小的速率上进行信号处理，例如自适应，那么系统设计的要求就降低了。该操作过程可以通过以下两种方式实现：一种就是时变的低通 FIR 滤波器可以设计工作在最原始的采样速率，$f_c=1/8$，其输出信号可以进行低速采样即每个采样间隔内减小 4 个点，即称为“低样抽取”。另外一种更有效的方式就是，滤波器可以分两步进行，即使用两次相同的简单单边带滤波器进行设计，每一次都采用较低的采样速率。这两种方案的示意图如图 12-7 所示。其中，第二种方案具有很明显的计算优势，这是由单边带滤波器的本质决定的；而且还可以将低样抽取单元转移到滤波器前面。根据前面介绍的调制过程，就可以采用同样的方式来实现低通和带通滤波器。

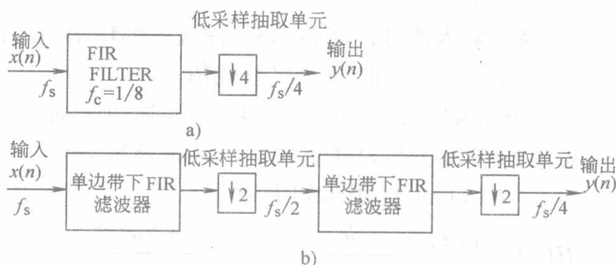


图 12-7　多速率 FIR 低通滤波器结构

a）方法 1　b）方法 2

自适应 FIR 滤波器的基本结构如图 12-8 所示。其中，自适应 FIR 滤波器的输入信号 $x(n)$ 用来为最后所需的 $d(n)$ 产生一个估算的 $\hat{d}(n)$；而 $e(n)$ 之间的差别通常被自适应算法用来控制 FIR 滤波器的加权系数。理想的响应信号主要和实际应用的需要相关。例如，在信道补偿中，如数据的可靠性通信，理想的响应信号就是接收端的一串时序信号。加权系数就会被进行调整以降低某些功能出现误差的几率，如均方值。最常见的自适应算法是最小平均二乘法，该算法减小了均方误差。这一类的滤波器可以适应时变环境，而固定滤波器则无法适应。

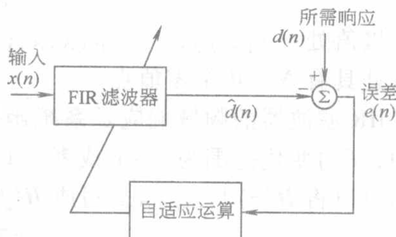


图 12-8　自适应 FIR 滤波器结构

5. 应用

数字 FIR 滤波器的精确性、再生的能力、多速率的实现以及适应时变环境的能力等方面的特性适合很多实际的应用，尤其是通信方面的应用，例如接收机和发射机的设计、语音压缩与编码以及信道复用等。固定加权系数 FIR 滤波器的主

要优点是指由于结构中无反馈而产生的无条件稳定特点以及它的精确相位特性。尽管如此，但是对于那些滤波要求严格并且有选择性滤波标准结构的应用来说，就需要高阶 FIR 滤波器；但是在某些应用中，高阶 FIR 滤波器可能是抑制滤波器，因此也是递归滤波器；而 IIR 滤波器也可以作为一个很好的选择。

12.3 IIR 滤波器

具有无限长脉冲响应的数字滤波器称为“无限脉冲响应（IIR）滤波器”。各种重要类型的 IIR 滤波器可以通过不同的公式来描述，如下所示：

$$y(n)=b_0x(n)+b_1x(n-1)+\cdots+b_Mx(n-M)-a_1y(n-1)-a_2y(n-2)-\cdots-a_Ny(n-N) \tag{12-1}$$

式中，$x(n)$ 是指滤波器的输入信号；$y(n)$ 是指滤波器的输出信号；$(a_1, a_2, \cdots, a_N)$ 和 $(b_0, b_1, \cdots, b_M)$ 是指实际的加权系数值。

我们用 $h(n)$ 表示脉冲响应，当 $n=0$ 且由单位脉冲驱动时，$h(n)$ 是系统的输出值，而此前系统的初始状态为静止状态。系统函数 $H(z)$ 是 $h(n)$ 的 z 域形式，根据式（12-1），$H(z)$ 可以表示如下：

$$H(z)=\frac{Y(z)}{X(z)}=\frac{b_0+b_1z^{-1}+\cdots+b_Mz^{-M}}{1+a_1z^{-1}+a_2z^{-2}+\cdots+a_Nz^{-N}} \tag{12-2}$$

式（12-2）中，N 称为滤波器的“阶”，式（12-2）可以转换成极值和零值的形式，如下所示：

$$H(z)=b_0z^{N-M}\frac{(z-q_1)(z-q_2)\cdots(z-q_M)}{(z-p_1)(z-p_2)\cdots(z-p_N)} \tag{12-3}$$

极值处于 $p_1, p_2, \cdots, p_N$ 点处；而零值处于 $q_1, q_2, \cdots, q_M$ 点处，同时在原点处具有 $N-M$ 个零值点。

IIR 滤波器的频域响应是系统函数在复平面单位圆上的估计值，即 $z=e^{j2\pi f}$。其中，f 的变化范围为 0 ~ 1 或者 −1/2 ~ 1/2。变量 f 表示数字频率。简单起见，我们可以将 $H(z)|_{z=\exp(j2\pi f)}$ 写成 $H(f)$，因此，$H(f)$ 可以表示如下：

$$H(f)=b_0e^{j2\pi(N-M)f}\frac{(e^{j2\pi f}-q_1)(e^{j2\pi f}-q_2)\cdots(e^{j2\pi f}-q_M)}{(e^{j2\pi f}-p_1)(e^{j2\pi f}-p_2)\cdots(e^{j2\pi f}-p_N)} \tag{12-4}$$

$$=|H(f)|e^{j\theta(f)} \tag{12-5}$$

式中，$|H(f)|$ 是指幅度响应；$\theta(f)$ 是指相位响应。

与 FIR 滤波器相比，在实现相同的幅度响应前提下，IIR 滤波器需要的阶数比 FIR 滤波器少很多。但是，FIR 滤波器通常比较稳定，而 IIR 滤波器的加权系数如果选择不合理就会出现不稳定。假设式（12-1）对应的是因果系统，那么如果 z 平面上的极值点全部处于单位圆内，该系统则是稳定的。由于稳定的因果 IIR 滤波器的相位无法达到线性，因此，在需要线性相位的应用中就必须选择

FIR 滤波器，而不能选择 IIR 滤波器。

1. 实现过程

式（12-1）给出了 IIR 滤波器的一种实现方式，具体结构如图 12-9a 所示，称为“直接形式Ⅰ”。通过对结构进行重新排列，我们可以得到“直接形式Ⅱ”，如图 12-9b 所示。通过移位，我们可以得到“移位直接形式Ⅰ”和“移位直接形式Ⅱ”，如图 12-9c 和图 12-9d 所示。

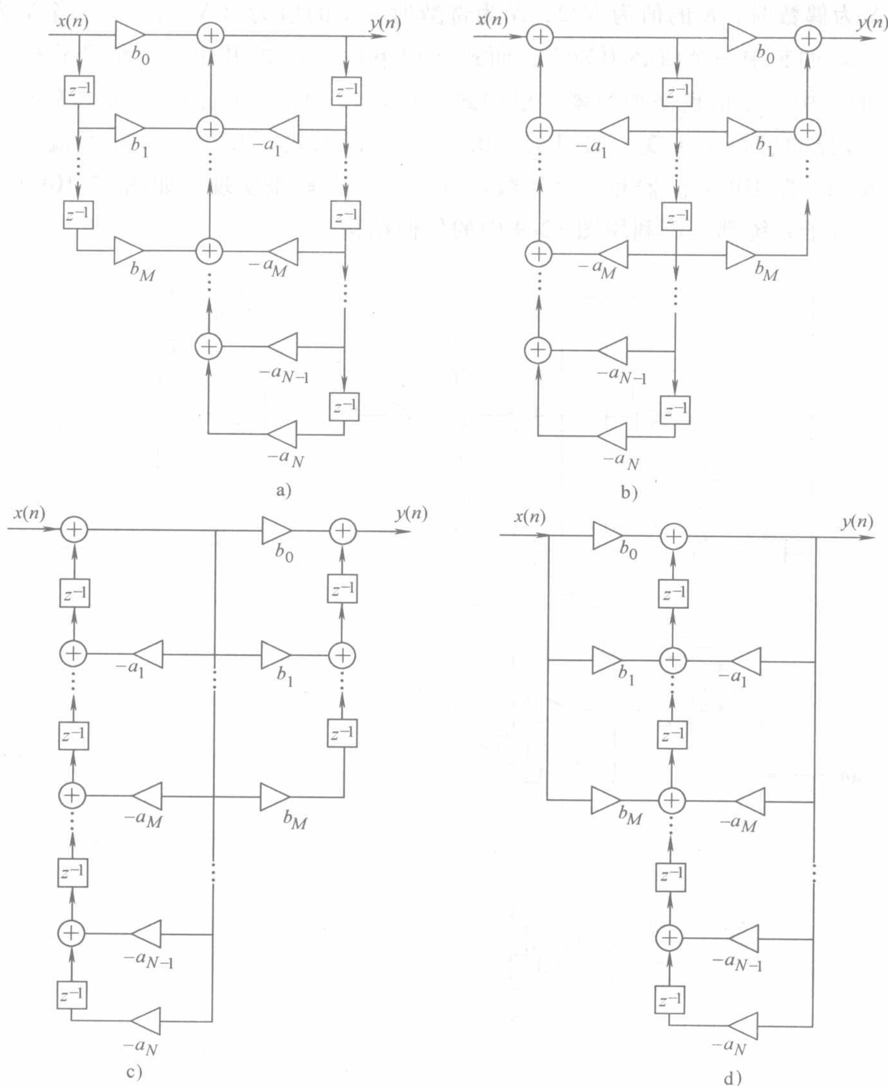


图 12-9　IIR 滤波器实现的直接形式

a）直接形式Ⅰ　b）直接形式Ⅱ　c）移位直接形式Ⅰ　d）移位直接形式Ⅱ

系统函数可以将分子和分母因式分解成二阶因子，转换成如下形式

$$H(z)=\prod_{i=0}^{K}\frac{b_{i0}+b_{i1}z^{-1}+b_{i2}z^{-2}}{1+a_{i1}z^{-1}+a_{i2}z^{-2}} \tag{12-6}$$

或者通过部分分数扩展，可以简化为

$$H(z)=c_0+\sum_{i=1}^{K}\frac{b_{i0}+b_{i1}z^{-1}}{1+a_{i1}z^{-1}+a_{i2}z^{-2}} \tag{12-7}$$

N 为偶数时，K 的值为 $N/2$；N 为奇数时，K 的值为（$N+1$）/2。当 N 为奇数时，a_{i2} 的其中一个值必须为零，而式（12-6）中 b_{i2} 的其中一个值和式(12-7)中 b_{i1} 的其中一个值也必须为零。根据式（12-6），IIR 滤波器可以通过 K 个二阶 IIR 滤波器的层叠来实现，如图 12-10a 所示。根据式（12-7），IIR 滤波器可以通过 K 个二阶 IIR 滤波器和一个计数器（如 c_0）并联来实现，如图 12-10b 所示。每个二阶子系统都可以利用图 12-9 中的任何结构。

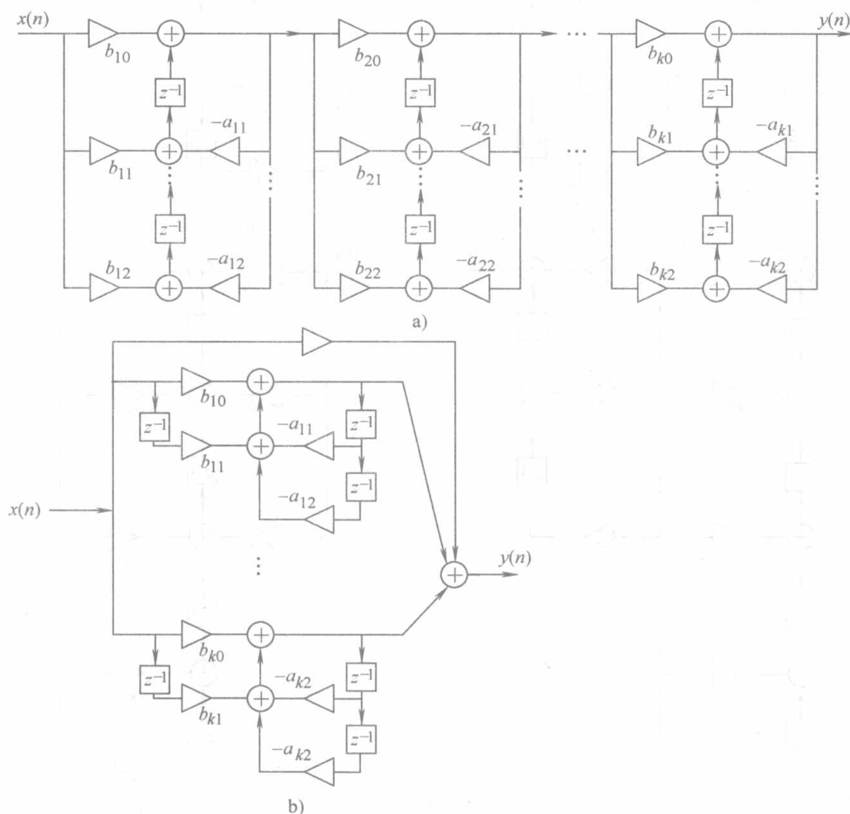


图 12-10 IIR 滤波器的实现方式

a）层叠形式 b）并联形式

IIR 滤波器还可以通过其他结构来实现，例如状态空间结构、波形结构以及晶格结构，具体过程读者可以参考相关的资料。在某些情况下，利用数字信号处理器的软件可以使 IIR 滤波器功能的实现变得更容易或者更方便。

2. IIR 滤波器的设计

IIR 滤波器的设计过程包含选取加权系数，以满足指定的要求，该要求通常是指幅度响应方面的规范。我们假设这样的要求形式如图 12-11 所示。其中，通频带幅度的平方必须在（$1/(1+\varepsilon^2)$，1）的范围内，而抑制频带幅度的平方不能超过 δ^2。通频带和抑制频带的边沿分别用 f_p 和 f_s 表示，而处于通频带和抑制频带中间的过渡频带不存在任何强制约束。

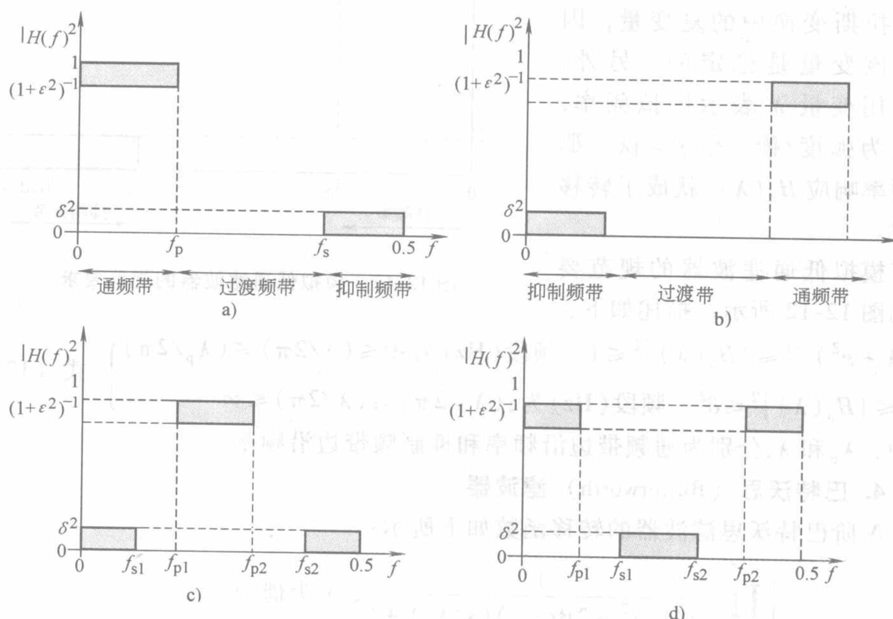


图 12-11　数字 IIR 滤波器的规范要求

a）低通滤波器　b）高通滤波器　c）带通滤波器　d）带阻滤波器

IIR 滤波器的设计方法各式各样：利用模拟原型滤波器进行设计、利用数字频率变换进行设计或者利用计算机辅助进行设计。在第一种方法中，模拟滤波器的设计必须满足（模拟）规范要求，而且模拟滤波器的转移函数变换成了数字系统函数；在第二种方法中，假设部分数字低通滤波器是有效的，而且理想滤波器是通过数字频率变换从数字低通滤波器得到的。在最后一种方法中，包含了加权系数的选择原则，以便各种响应尽可能地（某种程度上）接近理想的滤波器响应。前面两种方法很容易实现，而且它们非常适合标准滤波器（低通、高通、带通和带阻滤波器）的设计。虽然计算机辅助设计要求进行计算机编程，但是

这种方法可以用来设计标准的和非标准的滤波器。接下来我们将主要讨论第一种方法，并给出一些利用该方法进行设计的示例，同时还有对部分模拟滤波器的归纳总结。

3. 模拟滤波器

在此，我们归纳了3种基本类型的模拟低通滤波器，该滤波器可以作为IIR滤波器的设计原型。对于每一种类型，我们都给出了转移函数、幅度响应以及用来满足（模拟）规范要求的阶数N。我们用$H_a(s)$表示模拟滤波器的转移函数，其中s是拉普拉斯变换中的复变量，因此，该变量是稳定的。另外，我们用变量λ表示模拟频率，单位为弧度/秒。令$s=j\lambda$，那么频率响应$H_a(\lambda)$就成了转移函数。

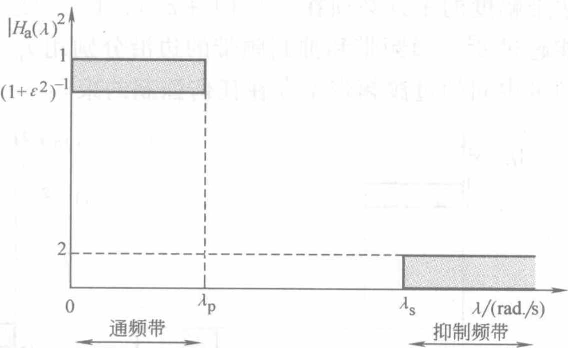


图 12-12　模拟低通滤波器的规范要求

模拟低通滤波器的规范要求如图12-12所示，描述如下：

$$\left.\begin{array}{ll}(1+\varepsilon^2)^{-1}\leqslant|H_a(\lambda)|^2\leqslant 1 & \text{频段(Hz)为}:0\leqslant(\lambda/2\pi)\leqslant(\lambda_p/2\pi)\\ 0\leqslant|H_a(\lambda)|^2\leqslant\delta^2 & \text{频段(Hz)为}:(\lambda_s/2\pi)\leqslant(\lambda/2\pi)\leqslant\infty\end{array}\right\}\quad\text{式 (12-8)}$$

式中，λ_p和λ_s分别为通频带边沿频率和抑制频带边沿频率。

4. 巴特沃思（Butterworth）滤波器

N阶巴特沃思滤波器的转移函数如下所示：

$$H_a(s)=\begin{cases}\prod_{i=1}^{N/2}\dfrac{1}{(s/\lambda_c)^2-2\mathrm{Re}(s_i)(s/\lambda_c)+1}, & N\text{为偶数}\\ \dfrac{1}{(s/\lambda_c)+1}\prod_{i=1}^{(N-1)/2}\dfrac{1}{(s/\lambda_c)^2-2\mathrm{Re}(s_i)(s/\lambda_c)+1}, & N\text{为奇数}\end{cases}\tag{12-9}$$

式中，$s_i=\exp\{j[1+(2i-1)/N]\pi/2\}$；$\lambda_c$为幅度下降3dB处对应的频率。

幅度响应的平方为

$$|H_a(\lambda)|^2=[1+(\lambda/\lambda_c)^{2N}]^{-1}\tag{12-10}$$

图12-13a给出了幅度响应$|H_a(\lambda)|$的示例。为了满足式（12-8），滤波器的阶数为

$$N=\text{integer}\geqslant\frac{\lg[\varepsilon/\delta^{-2}-1]^{\frac{1}{2}}}{\log[\lambda_p/\lambda_s]}\tag{12-11}$$

λ_c的值来自如下的范围：

$$\lambda_p \varepsilon^{-1/N} \leqslant \lambda_c \leqslant \lambda_s (\delta^{-2}-1)^{-1/(2N)} \tag{12-12}$$

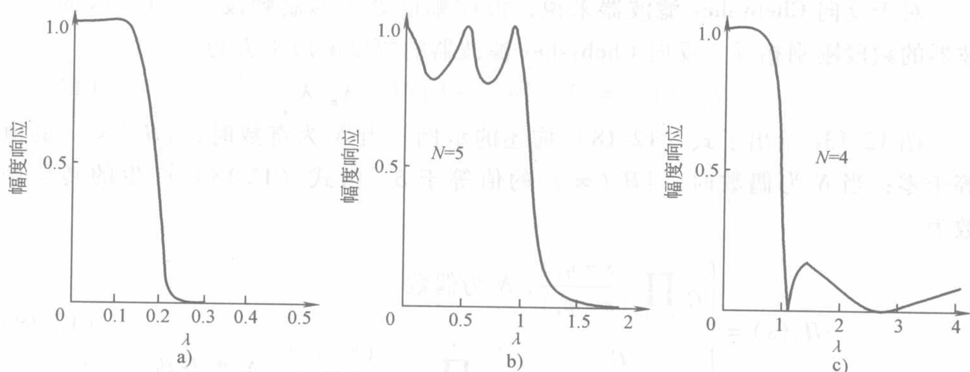


图 12-13 低通模拟滤波器的幅度响应

a）巴特沃思滤波器 b）Chebyshev 滤波器 c）反向 Chebyshev 滤波器

5. Chebyshev 滤波器（Ⅰ型 Chebyshev 滤波器）

N 阶 Chebyshev 滤波器的转移函数如下所示：

$$H_a(s) = C\prod_{i=1}^{N}\frac{1}{(s-p_i)} \tag{12-13}$$

式中，

$$p_i = -\lambda_p \sinh(\phi)\sin\left(\frac{2i-1}{2N}\pi\right) + \mathrm{j}\lambda_p \cosh(\phi)\cos\left(\frac{2i-1}{2N}\pi\right) \tag{12-14}$$

$$\phi = \frac{1}{N}\ln\left[\frac{1+(1+\varepsilon^2)^{\frac{1}{2}}}{\varepsilon}\right] \tag{12-15}$$

其中，当 N 为奇数时，$C = -\prod_{i=1}^{N} p_i$；N 为偶数时，$C = (1+\varepsilon^2)^{\frac{1}{2}}\prod_{i=1}^{N} p_i$。注意到，$C$ 对幅度进行了标准化，因此最大幅度为 1。幅度的平方可以表示为

$$|H_a(\lambda)|^2 = [1+\varepsilon^2 T_N^2(\lambda/\lambda_p)]^{-1} \tag{12-16}$$

上式中，$T_N(x)$ 是第一类递归公式中的 N 次 Chebyshev 多项式；该递归公式中，$T_0(x)=1$，$T_1(x)=x$，当 $n \geqslant 1$ 时，$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$。图 12-13b 给出了幅度响应平方的示例。注意到，在通频频段有相似的幅度波纹起伏。滤波器的阶数必须满足式（12-8），如下所示：

$$N \geqslant \frac{\lg\{[(\delta^{-2}-1)^{\frac{1}{2}}/\varepsilon] + [(\delta^{-2}-1)/\varepsilon^2 - 1]^{\frac{1}{2}}\}}{\lg\{(\lambda_s/\lambda_p) + [(\lambda_s/\lambda_p)^2 - 1]^{\frac{1}{2}}\}} \tag{12-17}$$

当 ε、δ、λ_p 和 λ_s 已知时，就可以计算出 N 的具体值。

6. 反向 Chebyshev 滤波器（Ⅱ型 Chebyshev 滤波器）

对于反向 Chebyshev 滤波器来说，波纹幅度处于抑制频段，和 Chebyshev 滤波器的频段刚刚相反。反向 Chebyshev 滤波器的幅度响应平方为

$$|H_a(\lambda)|^2 = [1 + (\delta^{-2} - 1)/T_N^2(\lambda_s/\lambda)]^{-1} \tag{12-18}$$

图 12-13c 给出了式（12-18）描述的示例。当 N 为奇数时，$|H_a(\infty)|$的值等于零；当 N 为偶数时，$|H_a(\infty)|$的值等于 δ。由式（12-18）产生的转移函数为

$$H_a(s) = \begin{cases} C\prod\limits_{i=1}^{N}\dfrac{(s-q_i)}{(s-p_i)},\ N\text{为偶数} \\ \dfrac{C}{(s-p_{(N+1)/2})}\prod\limits_{i=1,i\neq(N+1)/2}^{N}\dfrac{(s-q_i)}{(s-p_i)},\ N\text{为奇数} \end{cases} \tag{12-19}$$

式中

$$p_i = \frac{\lambda_s}{\alpha_i^2 + \beta_i^2}(\alpha_i - \mathrm{j}\beta_i);\ q_i = \mathrm{j}\frac{\lambda_s}{\cos\left(\dfrac{2i-1}{2N}\pi\right)} \tag{12-20}$$

$$\alpha_i = -\sinh(\phi)\sin\left(\frac{2i-1}{2N}\pi\right);\ \beta_i = -\cosh(\phi)\cos\left(\frac{2i-1}{2N}\pi\right) \tag{12-21}$$

$$\phi = \frac{1}{N}\cosh^{-1}(\delta^{-1}) = \frac{1}{N}\ln[\delta^{-1} + (\delta^{-2} - 1)^{\frac{1}{2}}] \tag{12-22}$$

其中，当 N 为奇数时，

$$C = \prod_{i=1}^{N}(p_i/q_i)$$

当 N 为偶数时，

$$C = -p_{(N+1)/2}\prod_{i=1,i\neq(N+1)/2}^{N}(p_i/q_i)$$

反向 Chebyshev 滤波器的阶数 N 必须满足式（12-8），如同式（12-17）定义的 Chebyshev 滤波器的阶数一样。

7. 比较

通过比较巴特沃思（Butterworth）滤波器和 Chebyshev 滤波器（包括反向 Chebyshev 滤波器），我们可以发现，在满足相同规范的条件下，巴特沃思滤波器需要的阶数比 Chebyshev 滤波器和反向 Chebyshev 滤波器都要高。还有一种滤波器称为“椭圆形”滤波器（Cauer 滤波器），该滤波器的通频带和抑制频带都存在波纹幅度。由于表达式太长，该类型的滤波器在此不作介绍（感兴趣的读者可以参考相关资料）。巴特沃思滤波器和反向 Chebyshev 滤波器在通频带具有比 Chebyshev 滤波器和“椭圆形”滤波器更好的相位特性（接近于理想特性），

但是在满足相同规范的条件下，“椭圆形”滤波器所需要的阶数比 Chebyshev 滤波器小。

8. 利用双线性变换进行设计

设计数字滤波器的一种最简单的方法就是通过对模拟低通滤波器进行变换，来实现理想的数字滤波器。从理想数字滤波器的规范要求出发，就可以得到低通模拟滤波器的规范。然后，设计模拟低通滤波器的 $H_a(s)$，以满足规范的要求。最后，通过将 $H_a(s)$ 变换成 $H(z)$ 就得到了理想数字滤波器的系统函数。这种变换包含 3 种类型，其中最全面、最综合的变换是双线性变换，本节将主要介绍该类型的变换。

在双线性变换过程中，$H_a(s)$ 的变量 s 被双线性函数 z 代替了，从而得到了 $H(z)$。表 12-1 给出了 4 种标准滤波器即低通滤波器（Low-Pass Filter，LPF）、高通滤波器（High Pass Filter，HPF）、带通滤波器（Band-Pass Filter，BPF）和带阻滤波器（Band-Stop Filter，BSF）的双线性变换的相关计算过程。表 12-1 中第二列给出了变量 s 和变量 z 之间的转换关系。T 的值可以任意选择，而且不会影响到最后的设计。第三列给出了模拟频率 λ 和数字频率 f 之间的转换关系，该转换公式是通过将变量 s 和变量 z 之间转换公式中的 s 用 $\mathrm{j}\lambda$ 代替，并将 z 用 $\exp(\mathrm{j}2\pi f)$ 来代替得到的。第四列和第五列给出了模拟 LPF 必需的通频带边沿频率和抑制频带边沿频率。注意到，模拟低通滤波器通频带和抑制频带允许的变化量或者说变量 ε 和 δ，与理想数字滤波器允许的变化量是相同的。另外注意到，对于 BPF 和 BSF 来说，变换过程是分两步进行的：第一步将模拟 LPF 变换成模拟 BPF（或者将 BSF 变换成 LPF）；第二步将模拟 BPF 变换成数字 BPF（或者将模拟 BSF 变换成数字 BSF）。W 和 $\tilde{\lambda}_0$ 的值由设计工程师自行设定。下面给出了几个选择：$W=\tilde{\lambda}_{p2}-\tilde{\lambda}_{p1}$，$\tilde{\lambda}_0^{\,2}=\tilde{\lambda}_{p1}\tilde{\lambda}_{p2}$，从而 $\lambda_p=1$；$W=\tilde{\lambda}_{s2}-\tilde{\lambda}_{s1}$，$\tilde{\lambda}_0^{\,2}=\tilde{\lambda}_{s1}\tilde{\lambda}_{s2}$，从而 $\lambda_s=1$。接下来，我们将通过两个示例来详细阐述该设计过程。

表 12-1　标准数字滤波器的双线性变换

	转换方式（s 与 z）	频率关系（f 与 λ）	模拟 LPF 的通频带边沿频率	模拟 LPF 抑制频带边沿频率
模拟 LPF 对应的数字 LPF	$s=\frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}$ $z=\frac{(2/T)+s}{(2/T)-s}$	$f=\frac{1}{\pi}\tan^{-1}\left(\frac{\lambda T}{2}\right)$ $\lambda=\frac{2}{T}\tan(\pi f)$	$\lambda_p=\frac{2}{T}\tan(\pi f_p)$	$\lambda_s=\frac{2}{T}\tan(\pi f_s)$
模拟 LPF 对应的数字 HPF	$s=\frac{T}{2}\frac{1+z^{-1}}{1-z^{-1}}$ $z=\frac{s+(T/2)}{s-(T/2)}$	$f=\begin{cases}-\frac{1}{2}+\frac{1}{\pi}\tan^{-1}\left(\frac{2\lambda}{T}\right),\lambda\geqslant 0\\ \frac{1}{2}+\frac{1}{\pi}\tan^{-1}\left(\frac{2\lambda}{T}\right),\lambda\leqslant 0\end{cases}$ $\lambda=\frac{T}{2}\tan[\pi(f+0.5)],\lambda\geqslant 0$	$\lambda_p=\frac{T}{2}\tan\left[\pi\left(f_p+\frac{1}{2}\right)\right]$	$\lambda_s=\frac{T}{2}\tan\left[\pi\left(f_s+\frac{1}{2}\right)\right]$

（续）

	转换方式（s 与 z）	频率关系（f 与 λ）	模拟 LPF 的通频带边沿频率	模拟 LPF 抑制频带边沿频率
模拟 LPF 对应的数字 BPF	$s=\frac{\tilde{s}^2+\tilde{\lambda}_0^2}{W\tilde{s}}$ 其中 $\tilde{s}=\frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}$	$\lambda=\frac{\tilde{\lambda}^2-\tilde{\lambda}_0^2}{W\tilde{\lambda}}$ 其中 $\tilde{\lambda}=\frac{2}{T}\tan(\pi f)$	$\lambda_p=\max\left\{\left\lvert\frac{\tilde{\lambda}_0^2-\tilde{\lambda}_{p1}^2}{W\tilde{\lambda}_{p1}}\right\rvert,\left\lvert\frac{\tilde{\lambda}_0^2-\tilde{\lambda}_{p2}^2}{W\tilde{\lambda}_{p2}}\right\rvert\right\}$ 其中 $\tilde{\lambda}_{p1}=\frac{2}{T}\tan(\pi f_{p1})$ $\tilde{\lambda}_{p2}=\frac{2}{T}\tan(\pi f_{p2})$	$\lambda_s=\min\left\{\left\lvert\frac{\tilde{\lambda}_0^2-\tilde{\lambda}_{s1}^2}{W\tilde{\lambda}_{s1}}\right\rvert,\left\lvert\frac{\tilde{\lambda}_0^2-\tilde{\lambda}_{s2}^2}{W\tilde{\lambda}_{s2}}\right\rvert\right\}$ 其中 $\tilde{\lambda}_{s1}=\frac{2}{T}\tan(\pi f_{s1})$ $\tilde{\lambda}_{s2}=\frac{2}{T}\tan(\pi f_{s2})$
模拟 LPF 对应的数字 BSF	$s=\frac{W\tilde{s}}{\tilde{s}^2+\tilde{\lambda}_0^2}$ 其中 $\tilde{s}=\frac{2}{T}\frac{1-z^{-1}}{1+z^{-1}}$	$\lambda=\frac{W\tilde{\lambda}}{\tilde{\lambda}_0^2-\tilde{\lambda}^2}$ 其中 $\tilde{\lambda}=\frac{2}{T}\tan(\pi f)$	$\lambda_p=\max\left\{\left\lvert\frac{W\tilde{\lambda}_{p1}}{\tilde{\lambda}_0^2-\tilde{\lambda}_{p1}^2}\right\rvert,\left\lvert\frac{W\tilde{\lambda}_{p2}}{\tilde{\lambda}_0^2-\tilde{\lambda}_{p2}^2}\right\rvert\right\}$ 其中 $\tilde{\lambda}_{p1}=\frac{2}{T}\tan(\pi f_{p1})$ $\tilde{\lambda}_{p2}=\frac{2}{T}\tan(\pi f_{p2})$	$\lambda_s=\min\left\{\left\lvert\frac{W\tilde{\lambda}_{s1}}{\tilde{\lambda}_0^2-\tilde{\lambda}_{s1}^2}\right\rvert,\left\lvert\frac{W\tilde{\lambda}_{s2}}{\tilde{\lambda}_0^2-\tilde{\lambda}_{s2}^2}\right\rvert\right\}$ 其中 $\tilde{\lambda}_{s1}=\frac{2}{T}\tan(\pi f_{s1})$ $\tilde{\lambda}_{s2}=\frac{2}{T}\tan(\pi f_{s2})$

9. 设计示例 1

假定设计一个数字 LPF，其通频带边沿频率为 $f_p=0.15$，抑制频带边沿频率为 $f_s=0.25$。通频带的幅度响应保持在 $-2\sim0$dB，而抑制频带的幅度响应不能超过 -40dB。假设模拟巴特沃思滤波器作为原型滤波器，那么设计过程如下所述。

1）计算 ε、δ、λ_p、λ_s和模拟滤波器的阶数。-2dB 的衰减意味着 $10\lg(1+\varepsilon^2)^{-1}=-2$，从而 $\varepsilon=0.7648$；-40dB 的衰减意味着 $10\lg(\delta^2)=-40$，从而 $\delta=0.01$。从表 12-1 中可以发现，模拟通频带和抑制频带边沿频率 $\lambda_p=(2/T)\tan(\pi f_p)$，而 $\lambda_s=(2/T)\tan(\pi f_s)$。为方便起见，我们令 $T=2$，这样可以得到 $\lambda_p=0.5095$，$\lambda_s=1.0$。因此，我们就可以根据式（12-11）来计算模拟巴特沃思滤波器所需的阶数了，计算结果为 $N\geqslant7.23$；通过取整，我们可以得到 $N=8$。

2）获取模拟 LPF 的转移函数。根据式（12-12），我们在 0.5269 和 0.5623 之间为 λ_c选取一个值。假设我们令 $\lambda_c=0.54$。加上之前得到的 $N=8$，根据式（12-9），我们可以得到转移函数，如下所示：

$$H_a(s)=\frac{7.2302\times10^{-3}}{(s^2+0.2107s+0.2916)(s^2+0.6s+0.2916)}\cdot\frac{1}{(s^2+0.8980s+0.2916)(s^2+1.0592s+0.2916)}\tag{12-23}$$

3）实现数字滤波器。利用表 12-1 中的变换，我们可以通过用 $(2/T)(z-1)/(z+1)=(z-1)/(z+1)$（因为 $T=2$）代替 s 来将式（12-23）变换成理想的数字滤波器公式。通过替代和简化，最后的数字 LPF 就具有了一个系统函数，如下所示：

$$H(z)=H_{\mathrm{a}}(s)\big|_{s=(z-1)(z+1)}=\frac{4.9428\times10^{-4}(z^2+2z+1)^4}{(z^2-0.9431z+0.7195)(z^2-0.7490z+0.3656)}\cdot\frac{1}{(z^2-0.6471z+0.1798)(z^2-0.6027z+0.0988)} \quad (12\text{-}24)$$

幅度响应 $|H(f)|$ 如图 12-14a 所示。通频带边沿频率 $f_{\mathrm{p}}=0.15$ 处的幅度为 0.8467，抑制频带边沿频率 $f_{\mathrm{s}}=0.25$ 处的幅度为 0.0072，这两个值都处于规范要求范围内。

10. 设计示例 2

在此，我们假定要设计一个数字 BSF，其通频带边沿频率分别为 $f_{\mathrm{p1}}=0.15$，$f_{\mathrm{p2}}=0.30$；抑制频带边沿频率分别为 $f_{\mathrm{s1}}=0.20$，$f_{\mathrm{s2}}=0.25$。通频带的幅度响应保持在 $-2\sim0\mathrm{dB}$，而抑制频带的幅度响应不能超过 $-40\mathrm{dB}$。我们采用 I 型 Chebyshev 滤波器作为原型滤波器。下面将介绍该类型的设计过程，该过程类似于上面示例 1 中的设计过程。

1）计算 ε、δ、λ_{p}、λ_{s} 和模拟滤波器的阶数。从示例 1 中，我们知道 $\varepsilon=0.7648$，$\delta=0.01$。根据表 12-1，并令 $T=2$，我们可以计算出模拟滤波器的通频带边沿频率和抑制频带边沿频率：$\tilde{\lambda}_{\mathrm{p1}}=\tan(\pi f_{\mathrm{p1}})=0.5095$，$\tilde{\lambda}_{\mathrm{p2}}=\tan(\pi f_{\mathrm{p2}})=1.7364$，$\tilde{\lambda}_{\mathrm{s1}}=\tan(\pi f_{\mathrm{s1}})=0.7265$，$\tilde{\lambda}_{\mathrm{s2}}=\tan(\pi f_{\mathrm{s2}})=1.0$。我们选择 $W=\tilde{\lambda}_{\mathrm{p2}}-\tilde{\lambda}_{\mathrm{p1}}=0.8669$，$\tilde{\lambda}_0^2=\tilde{\lambda}_{\mathrm{p1}}\tilde{\lambda}_{\mathrm{p2}}=0.7013$，因此 $\lambda_{\mathrm{p}}=1.0$，$\lambda_{\mathrm{s}}=\min\{3.6311,2.9021\}=2.9021$. 根据式（12-17），所需的模拟滤波器阶数 $N\geqslant3.22$，通过取整，我们可以得到 $N=4$。

2）获取模拟 LPF 的转移函数。根据式（12-13）和条件 $N=4$，我们可以得到模拟 Chebyshev 滤波器的转移函数，如下所示：

$$H_{\mathrm{a}}(s)=\frac{1.6344\times10^{-1}}{(s^2+0.2098s+0.9287)(s^2+0.5064s+0.2216)} \quad (12\text{-}25)$$

3）实现数字滤波器。利用表 12-1 中的变换，我们可以通过用 $W\tilde{s}/(\tilde{s}^2+\tilde{\lambda}_0^2)=W(z^2-1)/[(z-1)^2+\tilde{\lambda}_0^2(z+1)^2]$，（因为 $T=2$，$\tilde{s}=(z-1)/(z+1)$）代替 s 来将式（12-25）变换成理想的数字滤波器公式。通过替代和简化，最后的数字 BSF 就具有了一个系统函数，如下所示：

$$H(z)=\frac{1.7071\times10^{-1}(z^4-0.7023z^3+2.1233z^2-0.7023z+1)^4}{(z^4-0.5325z^3+1.1216z^2-0.4746z+0.8349)}$$

$$\frac{1}{(z^4-0.3331z^3+0.0660z^2-0.0879z+0.3019)} \tag{12-26}$$

幅度响应如图 12-14b 所示，该幅度响应满足规范要求。

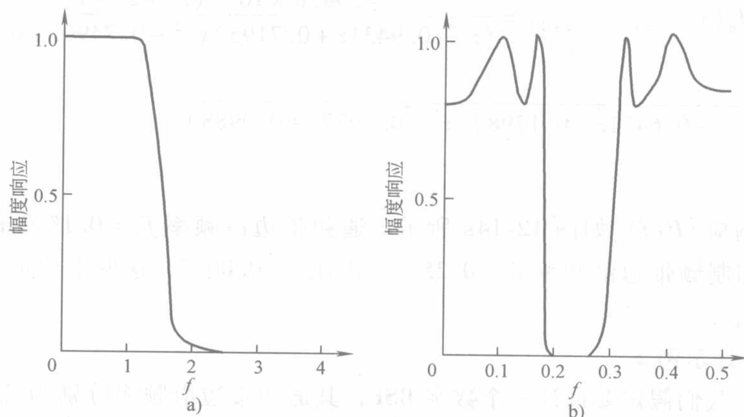


图 12-14 数字滤波器示例的幅度响应

a) LPF（示例 1） b) BSF（示例 2）

12.4 有限字长效应

实际中，数字滤波器在应用时必须同时配合有限精度和算术运算。因此，滤波器的加权系数和输入输出信号都是离散形式的，这就引出了 4 种类型的有限字长效应。

滤波器加权系数的离散化（量化）过程将会对滤波器极值点和零值点的位置产生影响。因此，实际的滤波器响应会和理想状态有轻微的差别。这种“决定性”的频率响应误差称为“系数量化误差”。

使用有限精度和算术运算时必须通过舍入或不舍入的方法来量化滤波器的计算过程。“舍入误差”是指在滤波器的输出信号中由于舍入或不舍入计算产生的误差。顾名思义，这种误差看上去类似于滤波器输出信号中的低电平噪声。

滤波器计算过程中的量化同样会导致滤波器出现轻微的非线性。对于强信号来说，这种非线性是可以忽略的，而主要关注的是舍入误差噪声。但是，对于只有一个零值点或输入信号为常量的递归滤波器来说，这种非线性会导致欺骗性的摆动，称为“极限环”。

如果采用定点算术运算，滤波器可能会产生计算溢出。“溢出摆动”有时也

称为“加法溢出极限环”，它是指存在于其他稳定的滤波器中的高电平摆动，这种摆动是由于与内部滤波器的计算溢出相关的非线性导致的。

本小节我们将对定点表示法和浮点表示法（数制）的这些有限字长效应进行研究。

1. 数值表示法

在数字信号处理过程中，$(B+1)$ 位的定点表示法表示的数值通常被描述成两个符号互补的形式，如：

$$b_0 b_{-1} b_{-2} \cdots b_{-B}$$

这样，该数值就变成了：

$$X = -b_0 + b_{-1}2^{-1} + b_{-2}2^{-2} + \cdots + b_{-B}2^{-B} \tag{12-27}$$

式中，b_0 是符号位，数值的范围为 $-1 \leqslant X < 1$。

这种表示法的优点是，在 $(-1, 1)$ 这个范围中两个数值的乘积仍然在该范围中。

浮点表示法描述的数值形式为

$$X = (-1)^s m2^c \tag{12-28}$$

式中，s 是符号位；m 是尾数；c 是（对数的）首数或指数。

为了使数值的表示法统一，尾数通常会被标准化，因此 $0.5 \leqslant m < 1$。

2. 定点量化误差

在定点运算中，乘法会使有效位的位数加倍。例如，两个 5 位的数值 0.0011 和 0.1001 的乘积就是 10 位的 00.00011011。小数点左边多余的位可以被丢弃，而不会产生任何误差。但是，剩下位中的最低 4 位最后必须通过某种量化形式丢弃，以便得到的结果可以保存为 5 位，供其他计算使用。在前面的示例中，这个结果是 0.0010（舍入量化）或者 0.0001（不舍入量化）。当进行乘积和运算时，量化过程可以在每个乘法完成之后进行，也可以在所有乘积进行双字长精度的求和之后再进行。

下面我们将对定点的舍入量化过程进行讨论。如果 X 是一个精确的数值，舍入的数值记为 $Q_r(X)$。如果量化值的小数点右边有 B 位，那么量化过程需要的步骤数为

$$\Delta = 2^{-B} \tag{12-29}$$

由于舍入过程选择的是最接近非量化值的量化值，因此舍入后的结果就不会超过精确值的 $\pm\Delta/2$。如果我们将舍入误差记为

$$\varepsilon_r = Q_r(X) - X \tag{12-30}$$

那么，

$$-\frac{\Delta}{2} \leqslant \varepsilon_r \leqslant \frac{\Delta}{2} \tag{12-31}$$

量化产生的误差可以进行建模，并作为一个随机变量均匀分布在相应的误差范围内。因此，具有舍入误差的计算过程可以看做是误差无关的过程，该过程会被白噪声中断。这种噪声的平均舍入值为

$$m_{\varepsilon_r}=E\{\varepsilon_r\}=\frac{1}{\Delta}\int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}}\varepsilon_r\mathrm{d}\varepsilon_r=0 \tag{12-32}$$

式中，$E\{\ \}$ 表示提取随机变量的有效值。

类似的，白噪声的舍入方差为

$$\sigma_{\varepsilon_r}^2=E\{(\varepsilon_r-m_{\varepsilon_r})^2\}=\frac{1}{\Delta}\int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}}(\varepsilon_r-m_{\varepsilon_r})^2\mathrm{d}\varepsilon_r=\frac{\Delta^2}{12} \tag{12-33}$$

3. 浮点量化误差

在浮点运算中，乘法和加法过程完成之后必须进行量化。必须进行加法量化（先于加法）的原因是由于在求和时小数值的尾数会右移直到进行加法的两个数值的指数（或幂）相同。通常，该过程会导致求和结果的尾数太长，因此需要进行量化。

我们假设浮点运算的量化过程通过舍入运算来完成。由于浮点运算中指数的原因，相对误差就成了最主要的了。相对误差的定义为

$$\varepsilon_r=\frac{Q_r(X)-X}{X}=\frac{\varepsilon_r}{X} \tag{12-34}$$

由于 $X=(-1)^s m2^c$，因此 $Q_r(X)=(-1)^s Q_r(m)2^c$

$$\varepsilon_r=\frac{Q_r(m)-m}{m}=\frac{\varepsilon}{m} \tag{12-35}$$

如果量化尾数的小数点右边具有 B 位，那么 $|\varepsilon|<\Delta/2$，如前所述，$\Delta=2^{-B}$，因此，加上 $0.5\leqslant m<1$，我们可以得到

$$|\varepsilon|_r<\Delta \tag{12-36}$$

如果我们假设 ε 均匀分布在 $-\Delta/2\sim\Delta/2$ 的范围内，而且 m 均匀分布在 $0.5\sim1$ 的范围内，那么

$$m_{\varepsilon_r}=E\left\{\frac{\varepsilon}{m}\right\}=0$$

$$\sigma_{\varepsilon_r}^2=E\left\{\left(\frac{\varepsilon}{m}\right)^2\right\}=\frac{2}{\Delta}\int_{\frac{1}{2}}^{1}\int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}}\frac{\varepsilon^2}{m^2}\mathrm{d}\varepsilon\mathrm{d}m=\frac{\Delta^2}{6}=(0.167)2^{-2B} \tag{12-37}$$

根据式（12-34），我们可以用非量化值和随机变量来描述量化后的浮点值，如下所示：

$$Q_r(X)=X(1+\varepsilon_r) \tag{12-38}$$

因此，有限精度乘积 X_1X_2 和求和 X_1+X_2 可以表示为

$$fl(X_1X_2)=X_1X_2(1+\varepsilon_r) \tag{12-39}$$

$$fl(X_1+X_2)=(X_1+X_2)(1+\varepsilon_r) \tag{12-40}$$

式中，ε_r 是指式（12-37）中变化量的零平均值。

4. 舍入噪声

为了确定数字滤波器输出端的舍入噪声，我们假设由于量化产生的噪声是固定的、白色的、与滤波器输入无关的内部变量。这个假设规避了滤波器的输入信号发生变化带来的复杂性。该假设不适用于零输入或常量输入，因为这两种情况下的舍入效应是从极限环的角度来分析的。

为了满足足够复杂输入信号的假设，数字滤波器的舍入噪声通常以零均值白噪声滤波器的输入信号 $x(n)$ 为原型来计算的，该输入信号的方差为 σ_x^2。这样就简化了输出舍入噪声的计算过程，因为 $k\neq0$ 时，$E\{x(n)x(n-k)\}$ 的有效值为零；当 $k=0$ 时，就可以计算出方差 σ_x^2。如果滤波器中还存在其他的舍入噪声源，那么我们就假设这些噪声之间是无关的，这样输出噪声的变化量就是所有噪声源的和。这种分析方式可以对输出舍入噪声进行有效估计，而且分析结果非常接近于在实际中观察到的值。

在计算舍入噪声时还有一种假设，就是两个量化误差的乘积为零。为了修正这种假设，我们讨论一个 16 位的定点处理器。在该处理器中，量化误差为 2^{-15} 级别，两个量化误差的乘积为 2^{-30} 级别，因此，比较而言，后者可以忽略不计。

最简单的分析例子就是通过卷积求和实现有限脉冲响应滤波器，如下所示：

$$y(n)=\sum_{k=0}^{N-1}h(k)x(n-k) \tag{12-41}$$

在定点运算中，量化过程是在乘法完成之后进行的，N 次乘法之后的量化噪声是单次乘法运算量化噪声的 N 倍。例如，根据式（12-29）和式（12-33），每次乘法运算之后的舍入运算会产生输出噪声差异，如下所示：

$$\sigma_0^2=N\frac{2^{-2B}}{12} \tag{12-42}$$

事实上，所有的数字信号处理器集成电路包含一个或多个双字长的累加寄存器，该寄存器可用于式（12-41）中乘积和的计算，而且无需量化。这种情况下，在求和之后只需要一个量化过程，而且方差为

$$\sigma_0^2=\frac{2^{-2B}}{12} \tag{12-43}$$

对于浮点舍入噪声的计算，我们假设式（12-41）中的 $N=4$，然后就可归纳出其他 N 值时的结果。有限精度输出可以用精确输出值加上一个误差值 $e(n)$ 来表示，如下所示：

$$\begin{aligned}y(n)+e(n)=(\{[&h(0)x(n)(1+\varepsilon_1(n))+h(1)x(n-1)(1+\varepsilon_2(n))][1+\varepsilon_3(n)]\\&+h(2)x(n-2)(1+\varepsilon_4(n))\}\{1+\varepsilon_5(n)\}+h(3)x(n-3)(1+\varepsilon_6(n)))(1+\varepsilon_7(n))\end{aligned} \tag{12-44}$$

在式（12-44）中，$\varepsilon_1(n)$ 是指第一个乘积的误差，$\varepsilon_2(n)$ 是指第二个乘积的误差；$\varepsilon_3(n)$ 是指第一个加法的误差，依此类推。必须注意，我们假设乘积是按照式（12-41）求和的顺序来求和的。

通过对式（12-44）进行展开并忽略各误差的乘积，那么可以得到：

$$e(n)=h(0)x(n)[(\varepsilon_1(n)+\varepsilon_3(n)+\varepsilon_5(n)+\varepsilon_7(n)]+h(1)x(n-1)[\varepsilon_2(n)+\varepsilon_3(n)+\varepsilon_5(n)+\varepsilon_7(n)]+h(2)x(n-2)[\varepsilon_4(n)+\varepsilon_5(n)+\varepsilon_7(n)]+h(3)x(n-3)[\varepsilon_6(n)+\varepsilon_7(n)] \tag{12-45}$$

假设输入信号是白噪声，其方差为 σ_x^2，因此，$k\neq 0$ 时，$E\{x(n)x(n-k)\}$ 的值为零，并假设误差之间是无关的，那么

$$E\{e^2(n)\}=[4h^2(0)+4h^2(1)+3h^2(2)+2h^2(3)]\sigma_x^2\sigma_{\varepsilon_r}^2 \tag{12-46}$$

通常来说，对于任何 N 值，

$$\sigma_0^2=E\{e^2(n)\}=[Nh^2(0)+\sum_{k=1}^{N-1}(N+1-k)h^2(k)]\sigma_x^2\sigma_{\varepsilon_r}^2 \tag{12-47}$$

注意到，如果卷积求和时乘积求和的顺序发生变化，那么式（12-47）中 $h(k)$ 出现的顺序就会发生变化。如果该顺序发生变化后最小幅度的 $h(k)$ 处于首位，紧接着是下一个最小幅度，依此类推，那么舍入噪声的变化就会达到最小。但是，卷积求和过程如果不是按照顺序进行的，那么就会使得数据索引变得非常复杂，而且使得舍入噪声的减小失去意义。

IIR 滤波器的舍入噪声分析方式和 FIR 滤波器是相同的。但是分析的过程相对更加复杂一些，因为在滤波器内部变量（状态变量）计算中产生的舍入噪声必须通过从量化点至滤波器输出端的一个转移函数来传播。这个过程对于通过卷积求和来实现的 FIR 滤波器不是必需的，因为所有的量化过程是在输出端的计算过程中进行的。采用定点运算实现 IIR 滤波器的另一个复杂地方是指必须对内部滤波器变量进行规划，以避免计算溢出。IIR 滤波器舍入噪声的分析示例可以参考 Weinstein 和 Oppenheim 在 1969 年以及 1972 年出版的著作，他们在著作中指出了，在滤波器实现结构之间的差异会导致输出端舍入噪声出现更大的差异。特别是，通过并联或层叠连接一阶和二阶子滤波器实现的 IIR 滤波器在舍入噪声性能方面通常优越于高阶直接形式（单差分方程式）实现的 IIR 滤波器。另外，在舍入噪声检测上可以选择最佳的或接近最佳的实现方式（参考 Mullins 和 Roberts 在 1976 年以及 Jackson，Lindgren 和 Kim 在 1979 年出版的著作）。这些实现方式通常需要更多的计算过程来根据输入样值获取输出样值。不过，舍入噪声稍大一点的次优实现方式也是可行的（参考 Bomar 在 1985 出版的著作）。

5. 极限环

极限环有时又称为"乘法器舍入极限环"，它是指一个低电平摆动，该摆动在其他稳定的滤波器中也存在，因此会产生和内部滤波器计算舍入（或非舍入）过程相关的非线性（参考 Parker 和 Hess 在 1977 年出版的著作）。极限环需要递归过程，在非递归 FIR 滤波器中不需要。

作为极限环的示例，我们假设二阶滤波器的实现函数为

$$y(n) = Q_r\left\{\frac{7}{8}y(n-1) - \frac{5}{8}y(n-2) + x(n)\right\} \tag{12-48}$$

式中，$Q_r\{\ \}$ 表示舍入的量化过程。

该稳定滤波器的极值点位于 0.4375 ± j0.6585。假设这种滤波器的实现过程中采用4 位（3 位和1 个符号位）的补码（二进制补码）定点运算，零初始条件为 $y(-1) = y(-2) = 0$，输入序列为 $x(n) = \frac{3}{8}\delta(n)$，其中 $\delta(n)$ 为单位脉冲或单位采样值。这样，就可以得到以下序列：

$$y(0) = Q_r\left\{\frac{3}{8}\right\} = \frac{3}{8} \qquad y(1) = Q_r\left\{\frac{21}{64}\right\} = \frac{3}{8} \qquad y(2) = Q_r\left\{\frac{3}{32}\right\} = \frac{1}{8}$$

$$y(3) = Q_r\left\{-\frac{1}{8}\right\} = -\frac{1}{8} \qquad y(4) = Q_r\left\{-\frac{3}{16}\right\} = -\frac{1}{8} \qquad y(5) = Q_r\left\{-\frac{1}{32}\right\} = 0$$

$$y(6) = Q_r\left\{\frac{5}{64}\right\} = \frac{1}{8} \qquad y(7) = Q_r\left\{\frac{7}{64}\right\} = \frac{1}{8} \qquad y(8) = Q_r\left\{\frac{1}{32}\right\} = 0$$

$$y(9) = Q_r\left\{-\frac{5}{64}\right\} = -\frac{1}{8} \qquad y(10) = Q_r\left\{-\frac{7}{64}\right\} = -\frac{1}{8} \qquad y(11) = Q_r\left\{-\frac{1}{32}\right\} = 0$$

$$y(12) = Q_r\left\{\frac{5}{64}\right\} = \frac{1}{8} \quad \cdots$$

注意到，尽管第一个采样的输入值为零，但是输出产生的摆动振幅为$\frac{1}{8}$，周期为 6。

定点递归滤波器主要关注的是极限环。当浮点滤波器采用并联或层叠连接一阶和二阶子滤波器来实现时，极限环通常就不会成为一个问题了，因为实际上极限环在一阶和二阶系统中是观察不到的，该一阶和二阶系统是通过 32 位的浮点运算实现的（Bauer，1993）。如前所述，这一类的系统必须具有一个非常低的稳定性冗余，以便极限环的等级不低于下溢的水平，该下溢水平是指振幅低于 10^{-38}。

当采用定点运算时，处理极限环的方法至少有 3 种。一种就是确定最大极限环幅度的范围，并作为量化的数字等级。然后，选择字长，该字长使得极限环幅度处于可接受的最低水平。第二种方法中，极限环可以由随机舍入计算来控制大

小（Buttner，1976），但是，这种方法很难实现。第三种方法就是合理选择滤波器的实现结构，并采用幅度不舍入方法对滤波器的计算过程进行量化（Bauer，1994），这种方法的缺点是会产生轻微的舍入噪声。

6. 溢出摆动

当采用定点运算时，滤波器的计算过程可能会产生溢出。只有当相同符号的两个数相加产生的结果数值振幅超过1时才会发生溢出。由于振幅比1大的数值无法表示，因此会发生溢出。例如，补码数0.101$\left(即\frac{5}{8}\right)$和0.100$\left(即\frac{4}{8}\right)$相加的结果为1.001，1.001就是$-\frac{7}{8}$的补码表达式。

补码表示法则的溢出特性可以记为“$R\{\ \}$”，如下所示：

$$R\{X\}=\begin{cases}X-2, & X\geqslant 1\\ X & -1\leqslant X<1\\ X+2, & X<-1\end{cases} \tag{12-49}$$

上面的例子中，$R\left\{\frac{9}{8}\right\}=-\frac{7}{8}$。

溢出摆动有时又称为“加法器溢出极限环”，它是指一个高电平摆动，该摆动在其他稳定的定点滤波器中也存在，它产生的主要原因是与内部滤波器计算的溢出（或非舍入）相关的非线性（Ebert，Mazo和Taylor，1969）。类似极限环，溢处摆动也需要递归过程，而在非递归FIR滤波器中不需要。

作为溢出摆动的示例，我们回过头再来关注一下式（12-48）对应的滤波器，该滤波器具有4位的定点补码运算法则，而且具有式（12-49）中的补码溢出特性，如下所示：

$$y(n)=Q_r\left\{R\left[\frac{7}{8}y(n-1)-\frac{5}{8}y(n-2)+x(n)\right]\right\} \tag{12-50}$$

在本例中，我们采用输入序列：

$$x(n)=\left\{-\frac{3}{4},-\frac{5}{8},0,0,0,\cdots\right\} \tag{12-51}$$

得到的输出序列为

$$y(0)=Q_r\left\{R\left\{-\frac{3}{4}\right\}\right\}=Q_r\left\{-\frac{3}{4}\right\}=-\frac{3}{4} \qquad y(1)=Q_r\left\{R\left\{-\frac{41}{32}\right\}\right\}=Q_r\left\{\frac{23}{32}\right\}=\frac{3}{4}$$

$$y(2)=Q_r\left\{R\left\{\frac{9}{8}\right\}\right\}=Q_r\left\{-\frac{7}{8}\right\}=-\frac{7}{8} \qquad y(3)=Q_r\left\{R\left\{-\frac{79}{64}\right\}\right\}=Q_r\left\{\frac{49}{64}\right\}=\frac{3}{4}$$

$$y(4)=Q_r\left\{R\left\{\frac{77}{64}\right\}\right\}=Q_r\left\{-\frac{51}{64}\right\}=-\frac{3}{4} \qquad y(5)=Q_r\left\{R\left\{-\frac{9}{8}\right\}\right\}=Q_r\left\{\frac{7}{8}\right\}=\frac{7}{8}$$

$$y(6)=Q_r\left\{R\left\{\frac{79}{64}\right\}\right\}=Q_r\left\{-\frac{49}{64}\right\}=-\frac{3}{4} \qquad y(7)=Q_r\left\{R\left\{-\frac{77}{64}\right\}\right\}=Q_r\left\{\frac{51}{64}\right\}=\frac{3}{4}$$

$$y(8)=Q_r\left\{R\left\{\frac{9}{8}\right\}\right\}=Q_r\left\{-\frac{7}{8}\right\}=-\frac{7}{8} \quad \cdots$$

这种大规模的摆动和实际幅度几乎完全一致。

在定点滤波器的实现过程中，有很多种方法可以有效阻止溢出摆动的发生。最明显的就是对滤波器的计算过程进行划分，使得溢出无法产生。但是，这种方法可能会无法有效限制滤波器的动态范围。另一种方法是强制使完整的乘积和达到饱和的 ±1，从而无法发生溢出（Ritzerfeld，1989）。仅仅使完整的乘积和达到饱和是很重要的，因为在二进制补码运算中间的溢出不会影响最后结果的精度。大多数定点数字信号处理器提供了完整乘积和的自动饱和功能，即触发“饱和运算”。另外，还有一种方法可以避免发生溢出摆动，就是利用滤波器的结构，使得内部滤波器的瞬变衰减至零（Mills，Mullis 和 Roberts，1978）。这种结构是很有效的，因为它们具有较低的舍入噪声，而且对加权系数量化过程不敏感（Barnes，1979）。

7. 加权系数量化误差

当滤波器的加权系数被量化为有限字长时，每一种滤波器结构就都具有了各自的有限个可实现的极值点和零值点分布。通常来说，滤波器中所需极值点和零值点的分布与可实现的分布并不完全一致。这种在滤波器实现过程中出现的误差（通常记为“频率响应误差”）称为“加权系数量化误差”，该误差主要来自于极值点和零值点的可实现位置与理想位置之间的差异。

假设二阶滤波器的复共轭极值点为

$$\lambda=re^{\pm j\theta}=\lambda_r\pm j\lambda_i=r\cos(\theta)\pm jr\sin(\theta) \tag{12-52}$$

转移函数为

$$H(z)=\frac{1}{1-2r\cos(\theta)z^{-1}+r^2z^{-2}} \tag{12-53}$$

实现过程中的差分方程式为

$$y(n)=2\lambda_r y(n-1)-r^2y(n-2)+x(n) \tag{12-54}$$

对差分方程式的加权系数进行量化，得到的结果是极值点的位置在 z 平面上呈现非均匀分布。非均匀分布是指垂直线的交叉部分对应的量化值为 $2\lambda_r$，而同心圆对应的量化值为 $-r^2$。接近 $z=\pm1$ 区域可实现的极值点非常少，因此在这些区域的极值就会产生大量的加权系数量化误差。相反，常见的实现过程中（Barnes，1979）量化加权系数对应了 λ_r 和 λ_i 的量化过程，这样最后可实现的极值点的位置就会呈现矩形网格状分布。

我们可以确定，具有低舍入噪声的滤波器的结构对加权系数的量化是兼容的，反之亦然（Jachson，1976；Rao，1986）。因此，常见的结构（矩形网格）由于它的低舍入噪声变得很受欢迎。

众所周知，多根高阶多项式根的分布是多项式系数中的一个非常敏感的函数。因此，如果高阶的滤波器是通过并联或层叠连接一阶和二阶子滤波器来实现的，那么滤波器的极值点和零值点就可以得到更加精确的控制。与此相关的就是线性相位 FIR 滤波器，该滤波器具有对称的多项式系数和单位圆附近的零值点，这样滤波器就可以通过卷积求和来直接实现。

8. 实现条件

线性相位 FIR 数字滤波器通常可以通过卷积求和的方式来实现，并且具有可接受的加权系数量化敏感性。如果在数字信号处理器上实现这种方式，那么定点运算不仅可以接受，而且比浮点运算更加合适。实际上，所有的数字信号处理器都是在双字长累加器中计算乘积和的，这就意味着只需要一个量化过程就可以计算出输出值了。另一方面，浮点运算在卷积求和的过程中每次乘法和加法完成之后都需要进行量化。在 32 位的浮点运算法则下，这些量化过程会产生一个足够小的误差，该误差对于大多数应用来说是无关紧要的。

在实现 IIR 滤波器时，任何通过并联或层叠连接一阶和二阶子滤波器的方式都比一个高阶直接形式的实现方式更适用。随着低功耗浮点数字信号处理器的使用，如 TI（Texas Instruments）公司的 TMS320C32 处理器，业界都推荐使用浮点运算法则来实现 IIR 滤波器。浮点运算同时还消除了大部分的规划、极限环和溢出摆动的顾虑。忽略采用的运算法则，多项式的二阶部分必须采用低舍入噪声的结构，采用这种结构通常会使滤波器具有较低的加权系数量化敏感性；而多项式的一阶部分对于实现过程中的舍入噪声和加权系数影响不大，因此该部分通常可以采用最直接的形式来实现。

名词解释

自适应 FIR 滤波器：有限脉冲响应结构滤波器，该滤波器具有可调整的加权系数。这种调整过程是通过自适应运算法则来控制的，如最小均方（Least Mean Square，LMS）法则，该法则广泛应用于通信系统的自适应回声消除器和补偿器中。

因果性：系统的一种特性，该特性表示系统的输出端只有在添加了输入信号后才会产生输出信号。这使得 FIR 滤波器在负值时间坐标上具有零离散时间响应。

离散时间脉冲响应：当输入信号为单位元素时，是 FIR 滤波器的输出值。

群时延：FIR 滤波器的群时延是指滤波器相位响应的负导数，因此也是输入频率的函数。在某个特殊频率，其值等于窄带信号通过滤波器的物理时延。

线性相位：FIR 滤波器的相位响应对于频率是线性的，因此，它也对应了常

量群时延。

线性时不变（Linear Time Invariant，LTI）："LTI 系统"是指该系统的输入值是由两个输入信号的和构成的，其对应的输出值也同样是由两个输出信号的和构成，这两个输出信号由两个输入信号分别独立产生，输出信号与输入信号的时间无关。

幅度响应：振幅的变化情况，在稳定状态下是通过 FIR 滤波器的一条正弦曲线，是频率的函数。

多速率 FIR 滤波器：一种 FIR 滤波器，其采样速率是可变的。

相位响应：相位的变化情况，在稳定状态下是通过 FIR 滤波器的一条正弦曲线，是频率的函数。

参 考 文 献

[1] Antoniou, A. 1993. *Digital Filters Analysis, Design, and Applications*, 2nd ed. McGraw-Hill, NewYork.

[2] Barnes, C. W. 1979. Roundoff noise and overflow in normal digital filters. *IEEE Trans. Circuits Syst.* CAS26 (3): 154-159.

[3] Bauer, P. H. 1993. Limit cycle bounds for floating-point implementations of second-order recursive digital filters. *IEEE Trans. Circuits Syst-Ⅱ*. 40 (8): 493-501.

[4] Bomar. B. W. 1985. New second-order state-space structure for realizing low roundoff noise digital filters. *IEEE Trans. Acoust., Speech, Signal Processing* ASSP-33 (1): 106-110.

[5] Bomar. B. W. 1994. Low-roundoff-noise linit-cycle-free implementation of recursive transfer functions on a fixed-point digital signal processor. *IEEE Trans. Industrial Electronics* 41 (1): 70-78.

[6] Buttner, M. 1976. A novel approach to eliminate limit cycles in digital filters with a minimum increase in the quantization noise. *In Processings of the* 1976 *IEEE International Symposium on Circuits and Systems*, pp. 291-294. IEEE, NY.

[7] Cappellini, V., Constantinides, A. G., and Emiliani, P. 1987. *Digital Filters and their Applications*. Academic Press, New York.

[8] Ebert, P. M., Mazo, J. E., anf Taylor, M. G. 1969. Overflow oscillations in digital filters. *Bell Syst. Tech. J.* 48 (9): 2990-3020.

[9] Gray, A. H. and Markel, J. D., 1973. Digital lattice and ladder filter synthesis. *IEEE Trans. Acoust., Speech, Signal Processing* ASSP-21: 491-500.

[10] Herrmann, O. and Schuessler, W. 1970. Design of nonrecursive digital filters with minimum phase. *Electronics Letters* 6 (11): 329-330.

[11] IEEE DSP Committee. 1979. *Programs for Digital Signal Processing*. IEEE Press, New York.

[12] Jackson, L. B. 1976. Roundoff noise bounds derived from coefficient sensitivities for digital filters. *IEEE Trans. Circuits Syst.* CAS-23 (8): 481-485.

[13] Jackson, L. B., Lindgren, A. G., and Kim, Y. 1979. Optimal synthesis of second-order state-space structures for digital filters. *IEEE Trans. Circuits Syst.* CAS-26 (3): 149-153.

[14] Lee, E. A. and, Messerschmitt, D. G. 1994. *Digital Communications*, 2nd ed. Kluwer, Norwell, MA.

[15] Macchi, O. 1995. *Adaptive Processing: The Least Mean Squares Approach with Application in Communica-*

tions, Wiley, New York.

[16] Mills, W. T., Mullis, C. T., and Roberts, R. A. 1987. Digital filter realizations without overflow oscillations. *IEEE Trans. Acoust.*, *Speech*, *Signal Processing* ASSP-26 (4): 334-338.

[17] Mullis, C. T. and Roberts, R. A. 1976. Synthesis of minimum roundoff noise fixed-point digital filters. *IEEE Trans. Circuits Syst.* CAS-23 (9): 551-562.

[18] Oppenheim, A. V. and Schafer, R. W. 1989. *Discrete-Time Signal Processing.* Prentice-Hall Englewood Cliffs, NJ.

[19] Oppenheim, A. V. and Weinstein, C. J. 1972. Effects of finite register length in digital filtering and the fast fourier transform. *Proc. IEEE* 60 (8): 957-976.

[20] Parker, S. R. and Hess, S. F. 1971. Limit-cycle oscillations in digital filters. *IEEE Trans. Circuits Theory* CT18 (11): 687-697.

[21] Parks, T. W. and Burrus, C. S. 1987. *Digital Filter Design.* Wiley, New York.

[22] Parks, T. W. and McClellan, J. H. 1972a. Chebyshev approximations for non recursive digital filters with linear phase. *IEEE Trans. Circuits Theory* CT-19: 189-194.

[23] Parks, T. W. and McClellan, J. H. 1972b. A program for the design of linear phase finite impulse response filters. *IEEE Trans. Audio Electroacoustics* AU-20 (3): 195-199.

[24] Proakis, J. G. and Manolakis, D. G. 1992. *Digital Signal Processing Principles*, *Algorithms*, *and Applications*, 2nd, et MacMillan, New York.

[25] Rabiner, L. R. and Gold, B. 1975. *Theory and Application of Digital Signal Processing.* Prentice-Hall Englewood Cliffs, NJ.

[26] Rabiner, L. R. and Schafer, R. W. 1978. *Digital Processing of Speech Signals.* Prentice-Hall Englewood Cliffs, NJ.

[27] Rabiner, L. R. and Schafer, R. W. 1974. On the behavior of minimax FIR digital Hilbert transformers. *Bell Sys. Tech. J.* 53 (2): 361-388.

[28] Rao, D. B. V. 1986. Analysis of coefficient quantization errors in state-space digital filters. *IEEE Trans. Acoust.*, *Speech*, *Signal Processing* ASSP-34 (1): 131-139.

[29] Ritzerfeld, J. H. F. 1989. A condition for the overflow stability of second-order digital filters that is satisfied by all scaled stated-space structures using saturation. *IEEE Trans. Circuits Syst.* CAS-36 (8): 1049-1057.

[30] Roberts, R. A. and Mullis, C. T. 1987. *Digital Signal Processing.* Addison-Wesley, Reading, MA.

[31] Vaiyanathan, P. P. 1993. *Multirate Systems and Filter Banks.* Prentice-Hall Englewood Cliffs, NJ.

[32] Weinstein, C. and Oppenheim, A. V. 1969. A comparison of roundoff noise in floating-point and fixed-point digital filter realization. *Proc. IEEE* 57 (6): 1181-1183.

备注

下面列出了一些关于数字滤波器的资料，感兴趣的读者可以参考相关的内容：

[1] 《*IEEE Transactions on Signal Processing*》，由电气和电子工程师学会出版的月刊，地址位于 345 East 47 Street，NY。

[2] 《*IEEE Transactions on Circuits and Systems-Part* Ⅱ：*Analog and Digital Signal Processing*》，由电气和电子工程师学会出版的月刊，地址位于 345 East 47 Street，NY。

电气和电子工程师学会每年都会在世界各地举办一次年会，称为“国际声学、语音和信号处理会议（International Conference on Acoustics，Speech and Signal Processing，ICASSP）”，总部地址位于 345 East 47

Street，NY。

[3]《*IEEE Transactions on Vision , Image and Signal Processing*》，由电气工程师学会出版的月刊，地址位于 Michael Faraday House，Six Hills Way，Stevenage，UK。

[4]《*Signal Processing*》，由欧洲信号处理协会出版，地址位于 Switzerland，Elsevier Science B. V.，Journals Dept.，P. O. Box 211，1000 AE Amsterdam，The Netherlands。

另外，下面所列的书籍也作为推荐资料：

[1] Bellanger，M. 1984. *Digital Processing of Signals: Theory and Practice.* Wiley，New York.

[2] Burrus，C. S. et al. 1994. *Computer-Based Exercises for Signal Processing Using MATLAB.* Prentice-Hall Englewood Cliffs，NJ.

[3] Jackson，L. B. 1986. *Digital Filters and Signal Processing.* Kluwer Academic，Norwell，MA，1986.

[4] Oppenheim，A. V. and Schafer，R. W. 1989. *Discrete-Time Signal Processing.* Prentice-Hall Englewood Cliffs，NJ.

[5] Widrow，B. and Stearns，S. 1985. *Adaptive Signal Processing.* Prentice-Hall Englewood Cliffs，NJ.

第 13 章　多片组件设计技术

Paul D. Franzon

13.1　引言

多片组件（MultiChip Module，MCM）设计技术可以将裸露的集成电路与有源器件一起布置在一个常见的互连衬底上，图 13-1 给出了一个多片组件设计技术的示例。在图 13-1 中，8 个芯片通过导线连接在一起构成了一个 MCM。不过，MCM 设计技术不仅介绍了芯片间的封装技术，还可以帮助我们挖掘芯片在容量和性能以及成本方面的优势。本章的主要内容包括：介绍各种不同 MCM 设计技术、阐述 MCM 如何推动系统性价比、说明 MCM 系统设计中需要注意的事项。

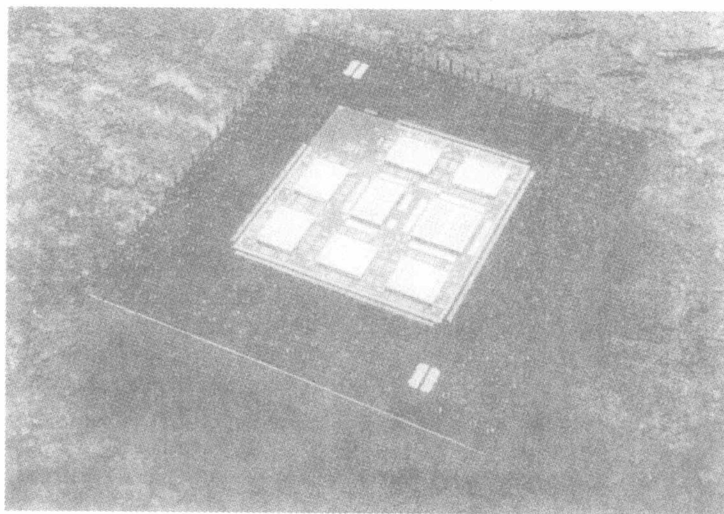


图 13-1　8 个芯片通过导线连接在一起构成的 MCM（来源：MicroModule Systems）

13.2　多片组件设计技术的定义

从广义上来说，多片组件设计技术是指将多个集成电路（IC）聚集在一个

常见的互连衬底上。这个定义看起来有些狭隘，它给我们的感觉就是常见的衬底比传统的印制电路板（PCB）提供了更高的连线集成度。MCM 的主要组成器件如图 13-2 所示，具体描述如下：

1）衬底提供了芯片间以及芯片与其他独立电路元件之间的互连互通功能，这些独立元件包括：电阻、电容和电感。

2）芯片互连技术为衬底和芯片之间信号和电源的传输提供了通道。

3）MCM 封装技术是指对 MCM 进行封装，并为 MCM 中的信号、电源和热量提供与外界交流的通道。

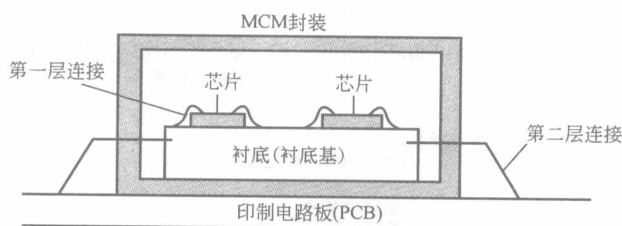


图 13-2　MCM 的物理构造

图 13-2 中没有说明的其他 MCM 重要部分包括：

1）测试技术，用来确保对裸芯片、MCM 衬底以及组装好的 MCM 进行校正。

2）维修技术，用来替换组装后检测到的失效冲模。

3）设计技术。

MCM 技术包含了各种不同的技术；其中，衬底技术主要可以划分为以下 3 种类型。

1. 碾压 MCM 技术

碾压 MCM（Laminate MCM，MCM-L）技术如图 13-3 所示。实质上，精细线路 PCB 和 MCM-L 是由第一层玻璃纤维片/填充树脂片上的铜制导片构成的，如图 13-4 所示。这些纤维片/树脂片在高温和高压条件下被碾压在一起。不同纤维片/树脂片上导片之间的互连是通过电镀的通道钻孔来实现的。MCM-L 技术的最新发展重点已经转移到了活动碾压技术上面。相比基于玻璃纤维的碾压技术，活动碾压技术可以实现更好的布线质量和通孔质量。

2. 陶瓷 MCM 技术

陶瓷 MCM（Ceramic MCM，MCM-C）技术主要基于引脚网格阵列（Pin Grid Array，PGA）技术。图 13-5 给出了 MCM-C 横截面和几何尺寸，图 13-6 列出了 MCM-C 的基本制造步骤。MCM-C 的制造首先通过铸造一片统一的预先焙烧陶瓷

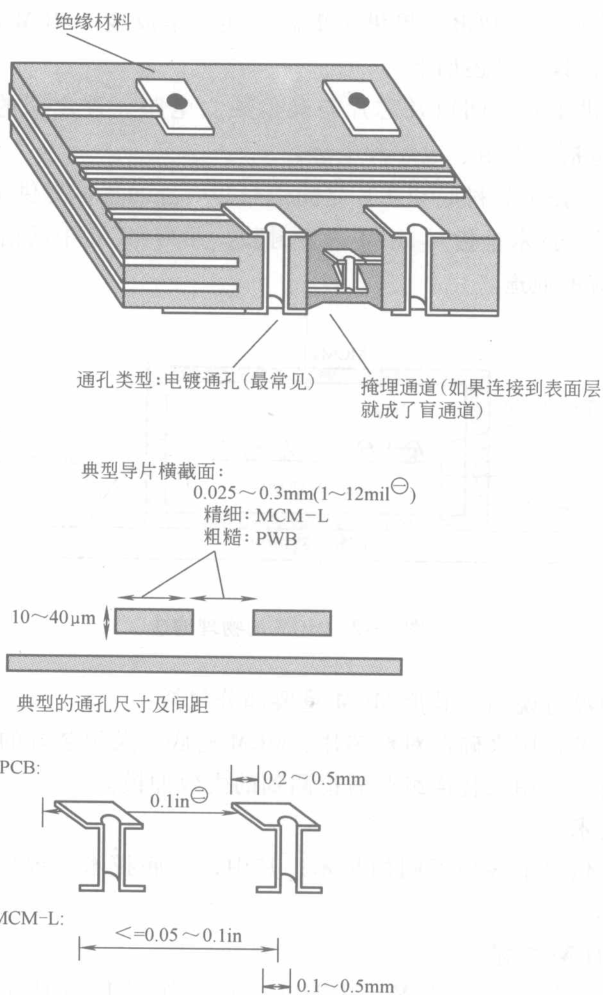


图 13-3 印制电路板上的典型横截面和特征尺寸以及 MCM-L 技术

材料，称为“未加工带”；然后在未加工带上印制金属涂料；再进行钻孔和电镀；最后将层叠的陶瓷片在一定压力下烘烤在一起。除了金属涂料之外，也可以

㊀ 1mil = 25.4 × 10^{-6}m。

㊁ 1in = 0.0254m。

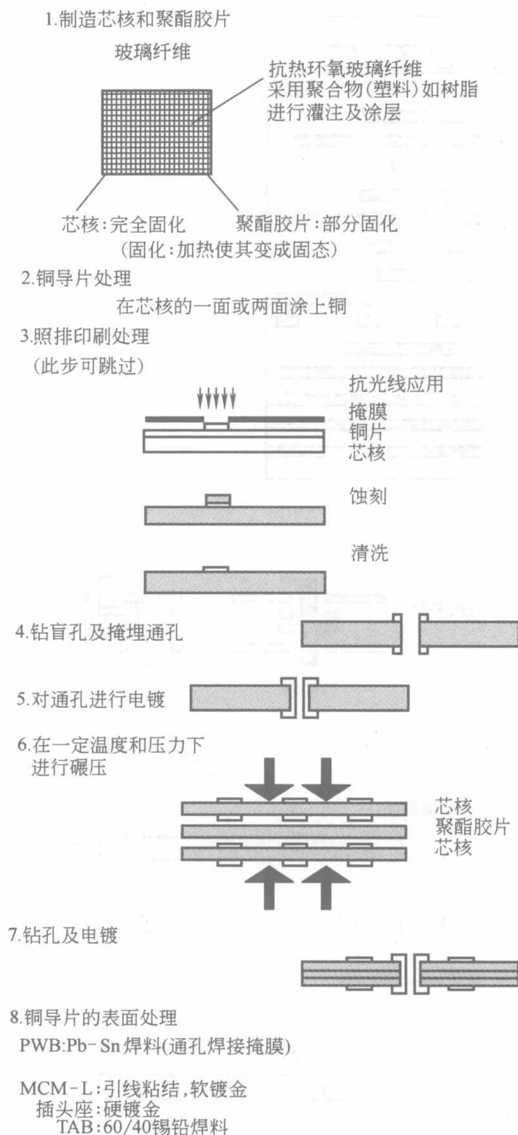


图 13-4　MCM-L 的基本制造步骤（采用玻璃纤维片/填充树脂片作为基座）

使用其他涂料，包括用于印制 MCM-C 中独立电阻和电容的涂料。

3. 薄膜式（沉积）MCM 技术

薄膜式（沉积）（Deposited MCM，MCM-D）技术主要基于芯片金属化处理

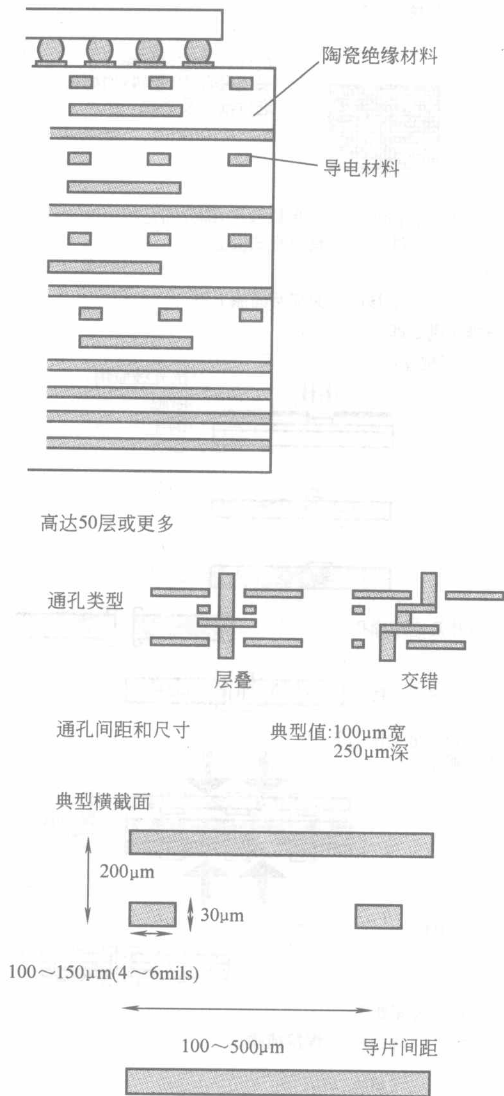


图 13-5　MCM-C 技术中的典型横截面和特征尺寸

技术。MCM-D 具有很好的特征尺寸，可以实现很高的布线集成度（见图 13-7）。MCM-D 每次只能制造一层，主要是根据金属导片模式、钻孔和绝缘层沉积物（通常为聚酰亚胺，如图 13-8 所示）的光刻精确度来实现的。通常，MCM-D 是建立在硅衬底上的，该衬底还可以承载电容、电阻和晶体管作为衬底的一部分，而且成本低廉。

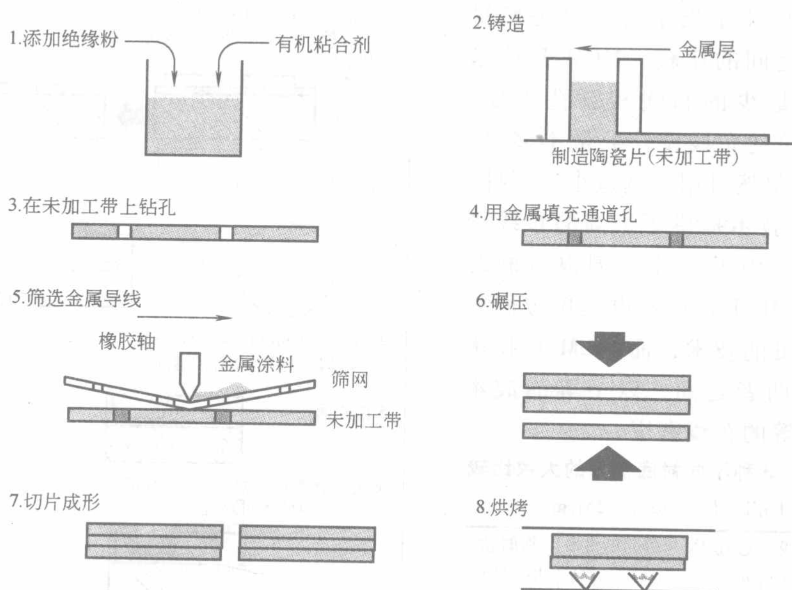


图 13-6　MCM-C 的基本制造步骤

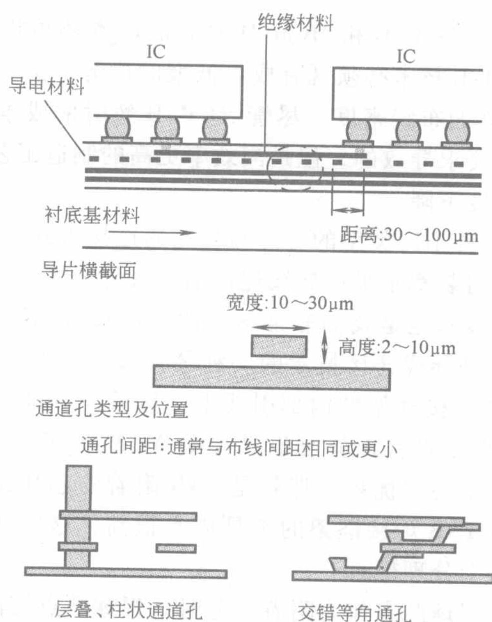


图 13-7　MCM-D 技术中的典型横截面和特征尺寸

表 13-1 给出了各种可替换衬底技术之间的比较。MCM-L 技术提供了最少的布线和钻孔密度，但是当总的布线设计数量不多时，MCM-L 仍然可用于制造小型器件；MCM-D 技术提供了最高的布线密度，主要用于设计引脚很多的芯片，但 MCM-D 技术也是单位面积上最昂贵的技术；而 MCM-C 技术处于这两者之间，以中等的成本提供中等的布线密度。

表 13-1 3 种不同衬底技术的大致比较

(1mil = 1/1000in ≈ 25μm)

	最小导线间距/μm	通孔大小/μm	每部分的价格/(美元/in²)	临时价格/美元
PCB	300	500	0.3	100
MCM-L	150	200	4	100 ~ 10000
MCM-C	200	100	12	25000
MCM-D	30	25	20	15000

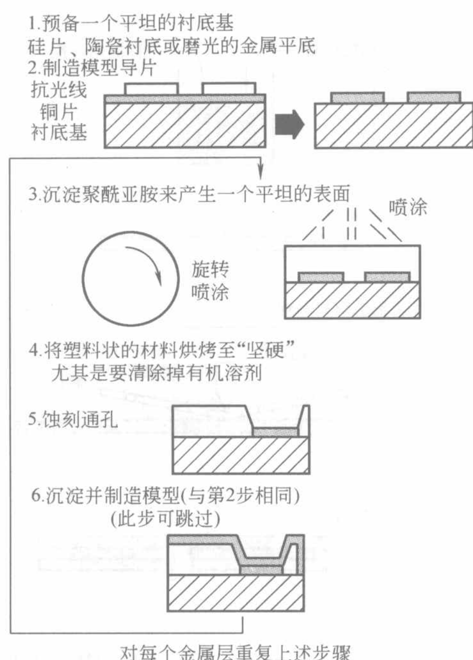


图 13-8 MCM-D 的基本制造步骤

最新的研究表明，MCM-L 和 MCM-D 衬底的成本还可以进一步下降。基于活动衬底技术的 MCM-L 技术必须既有成本低廉的优势，又可以提供比现有玻璃纤维 MCM-L 技术更高的布线密度。尽管 MCM-D 暂时的成本仍然很高，这主要是因为对光刻模板的要求导致的，但是当效率更高的制造工艺实现之后，每个元器件的成本将会大幅度下降。

图 13-9 给出了几种不同类型的芯片与衬底的连接方式。目前，MCM 中超过 95% 的冲模或单芯片封装都是采用导线连接的。大部分裸芯片非常适用导线进行连接，而且导线连接技术也是我们非常熟悉的。带状自动压焊（Tape Automated Bonding，TAB）技术是导线连接技术的一种替代技术。如果采用 TAB 技术，芯片首先必须贴在 TAB 焊接框架的内部引线上。然后，对这些引线设计形状（称为“定形”），之后外部的引线就可以粘接到 MCM 上了。相比导线连接技术，TAB 技术具有一个很重要的优势，那就是 TAB 附着的芯片比裸芯片更加容易进行测试。但是，制造 TAB 焊接框架的工具成本很高，这就使得该技术在芯片的大批量生产方面不是十分理想。

当采用倒装晶片焊球凸点作为附着部分时，焊球凸点沉淀在晶片的整个区域上。然后，晶片被切割成各个单独的冲模，并附着在 MCM 衬底上。之后，上述过程形成的模块被放进一个回流烘炉里面。在烘炉中，焊料就会在芯片和沉淀之

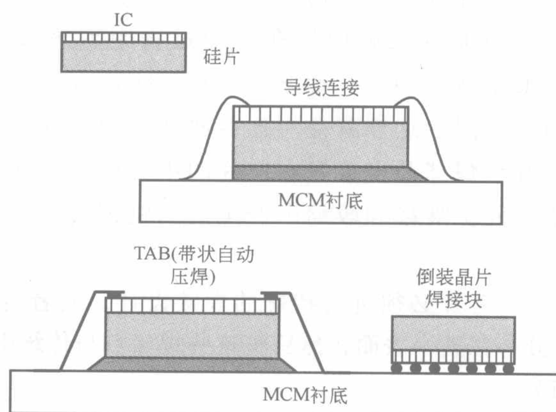


图 13-9　芯片与衬底的连接选择

间形成很强的粘合力。倒装晶片附着部分具有以下优势：

1）焊球凸点可以被放置在芯片的整个区域上，并允许芯片拥有数千个连接点。例如，一个 $1cm^2$ 的芯片可以支持 1600 个焊球凸点（保守间距为 250μm），但是只支持 533 个导线连接点。

2）晶体管可以被放置在焊球凸点下面，但是不能放置在导线连接点或 TAB 焊点下面，其原因和相关的附着部分处理过程有关。在制造导线连接或 TAB 引线连接部分时，需要将粘结剂压进芯片焊点里面，但同时这个过程会毁坏焊点下面的晶体管。另一方面，良好的焊接只能通过加热的方式实现。最后，芯片的尺寸可以减小相当于整个焊接环的面积。例如，一个 $100mm^2$ 的芯片，其焊接环的直径为 250μm。焊接环占用的面积总共可达到 $10mm^2$，约为芯片面积的 10%；如果芯片采用倒装和焊球凸点技术的话，那么这块面积就可以省下来，以供其他功能元器件使用。因此，芯片的尺寸就可以变得更小，成本就可以降得更低。

3）焊球凸点技术产生的电气寄生效应远比导线连接或 TAB 引线产生的寄生效应好。导线连接或 TAB 引线通常会在电路中产生约 1nH 的电感和 1pF 的电容。而焊球凸点产生约 10pH 的电感和 10nF 的电容。更理想的寄生电感和电容使得焊球凸点技术更受高频无线电应用的欢迎，例如，在 5.6GHz 的通信频段和高时钟速率数字器件中的应用。

4）倒装晶片的成本和导线连接技术的成本差不多，略微高一点。

但是，倒装晶片焊球凸点也存在一些缺陷。其中，最主要的就是倒装晶片焊接点比导线连接的焊接点大（分别为 60 ~ 80μm 和 50μm），这是由其规范

要求决定的，倒装晶片焊接点必须稍大一些，以便使焊球凸点更高一些。更高的焊球凸点可以很容易地缓解模块在装配过程中加热或变冷时产生的压力。由于硅的热扩展系数（TCE）通常和MCM衬底的不同，因此这两种材料的扩展率和收缩率不同，这样就会对连接两部分的焊球凸点产生压力。基于这样的原因，最好将焊球凸点放置在芯片的中心位置，而不要放在邻近边缘位置。通过减小不同扩展率和收缩率产生的间距，就可以减小焊球凸点的压力。

更大的焊接点要求芯片必须进行特殊的焊接设计，或者对晶片进行后期加工，以使焊球凸点分布在整个表面，然后将这些焊接点与传统引线焊接点进行导线连接（重新分布）。

另外一个潜在隐患是焊球凸点中铅的含量。大多数的铅中都包含有放射性的同位素，它是阿尔法粒子产生的源头，阿尔法粒子会潜在地改变邻近没有铝片保护的晶体管的状态。阿尔法粒子主要会对动态随机存取存储器（DRAM）和动态逻辑单元产生影响。

13.3 设计、修补和测试

物理技术中一个重要的补充技术就是对掩模和模块的测试和修正技术。每一个MCM都会涉及到一个重要的问题，即在安装掩模之前对其进行测试的成本是多少，以及依靠安装之后的测试来定位不合格冲模的成本又是多少。这完全是一个成本的问题。裸芯片测试的次数越多，安装后MCM正常工作的可能性越大。如果安装后MCM无法正常工作，那么该MCM必须进行修补（换一个冲模），或者干脆被丢弃。这样就只剩下一个问题了，那就是裸芯片的测试要达到什么样的程度才能使安装冲模后的MCM可以确保正常工作；或者说安装模块的成本便宜到什么程度，才能任意丢弃有缺陷的冲模。

通常，裸芯片的测试和预烧具有4个等级，即所谓的“知名优质冲模（Known Good Die，KGD）”的4个等级。表13-2给出了关于这4个测试等级的概述及其影响。其中，最低等级的KGD测试是指与晶片等级相同的测试，称为“晶片类型测试”。在此，芯片必须进行低速的功能性测试，同时还有参数测定（例如，晶体管特性曲线）。在最低等级的KGD测试中，裸芯片的测试成本必须限制在晶片测试的成本之下。但是，该测试过程仍然存在一定风险，即当芯片作为MCM的一部分被测试时，芯片是无法工作的。这种风险称为“测试中断率”。在传统的封装过程中，芯片封装后还要进行一次测试，以确保测试中断率为零。

表 13-2　4 个 KGD 测试等级及其影响

KGD 等级	测试成本效率	测试遗漏产生的影响
晶片等级功能性和参数测试	低	成熟 IC 的遗漏率约 <1% ~ 2%；新的 IC 遗漏率约 >5%
指定的引脚速度下的测试	中等	对于新的 IC，遗漏率最小
预烧条件下的测试	高	对于存储器件，预烧非常重要
全面测试条件下的动态预烧	最高	对于存储器件，遗漏率最小

尽管如此，如果 MCM 包含的芯片必须满足严格的时间要求（如工作站），或者 MCM 必须满足很高的可靠性标准，那么就需要进行更高等级的 KGD 测试。例如，工作站就希望他的芯片达到指定的速度。速度分类就要求提供专门的、能在测试器之间运行高速信号的测试装置。用于测试晶片种类的测试装置通常称为"探针板"，该装置不能运行高速信号。因此，必须重新构建一个更加昂贵的高速装置，或者将裸芯片安装在一个临时的测试包中。因此，这些额外的开销增加了引脚指定速度冲模的成本。

另外，在有些应用中需要等级更高的 KGD。例如，航天设备通常就需要预烧的冲模（尤其是存储器），以便降低初次失败的几率。这种预烧的冲模包含两种测试等级，其中级别较低的冲模测试是在一段时间的高温和高压条件下进行的；而在级别较高的测试中，冲模是烘炉中进行的。

如何确定哪个测试等级适合你的 MCM 呢？这个问题的答案由成本决定，即测试成本和维修成本。例如，假设一个 4 芯片的 MCM 采用了成熟的 IC，而成熟的 IC 具有非常高的效率，而且处理工程师也非常熟悉如何防止出现差错，那么成熟 IC 通过晶片级功能性测试的可能性就非常高。如果测试中断率为 2%，那么每个 MCM 在安装之后需要再更换芯片的可能性就只有 8%（$1-0.98^4$）。如果每次更换芯片的成本为 30 美元，那么每个 MCM 额外的平均维修成本约为 2.40 美元。更高级别的 KGD 测试成本一般比每个芯片的总体成本高至少 0.6 美元，因此，一般采用级别较低的测试。

另一方面，假设 MCM 中包含 4 个粗糙的 IC，那么功能性和参数晶片级测试就很难捕捉到速度缺陷。另外，处理工程师也没有机会来了解如何使芯片的效率最大以及如何检测潜在的问题了。如果测试中断率为 5%，那么 4 芯片的 MCM 需要更换芯片的几率就变成了 40%；这种情况下，每个 MCM 需要的平均维修成本就变成了 12 美元，而且还需要额外增加关于获取指定速度冲模的测试成本。

对于高速系统来说，速度分类会严重影响模块的速度等级。例如，当前，微处理器就是根据它们的时钟速率来分级的。速度越快的元器件，其价格越高；在 MCM 中，整个模块的运行速度直接和运行最慢的芯片相关，而且当一个系统被划分为多个更小的模块后，如果不经过测试，每个 MCM 中就很可能包含速度较

慢的芯片。

例如，假设一组芯片在制造时包含相同数量的较慢的、中等的和较快的芯片，如果芯片在安装到 MCM 之前根据速度进行分类，那么就有约 33% 的速度较慢的模块、33% 的速度中等的模块以及速度较快的模块。如果不根据速度分类，这些不同速度的模块的比率就会发生戏剧性地变化，如图 13-10 所示。对于一个没有事先对冲模进行速度分类的 4 芯片模块来说，将会产生 80% 的慢速率系统、19% 的中速率系统，而快速率系统只占剩下的 1%。快速率系统比率的大幅度降低证明了事先对冲模进行速度分类是很有必要的。

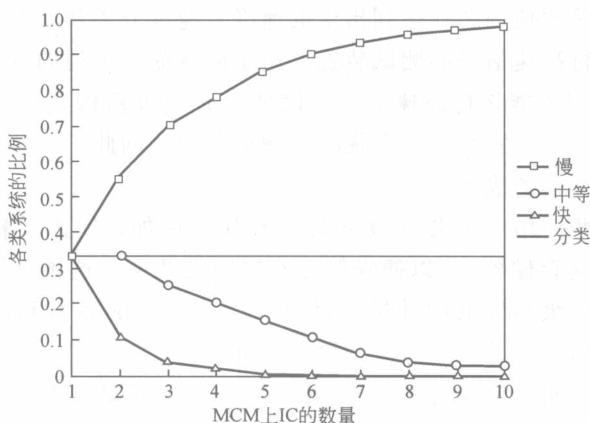


图 13-10 速度分类对 MCM 性能的影响

在 MCM 中，设计技术也是一个很重要的组成部分。大多数 MCM 计算机辅助设计（CAD）工具是 PCB 设计工具的基础，两者之间的主要区别是其内在的特征内容，这些内在特征包括允许在 MCM 中广泛应用各种物理几何学（尤其是通道几何学），以及可以使用裸芯片。一种新的版式工具（称为“冲模互换版式”）已经被开发出来，专门用于处理与裸芯片相关的物理信息（如焊点位置）。

除了对物理设计工具进行改进之外（那些工具实际上只能确定 MCM 的布线模式），MCM 的设计技术还有很多工作要做。MCM 中设计的正确性远比 PCB 中的正确性重要。例如，用来纠正错误的跳线在 MCM 中是很难布置的。因此，设计工具的最新发展已经集中到了对设计师能力的提升上来了，这样才能确保多芯片系统设计在构建之前的正确性。设计工具的最新发展包括新的仿真库和工具以及可以自动进行 MCM 电气设计和热学设计的工具；其中仿真库使设计师可以在构建 MCM 之前对整个芯片进行仿真。

13.4 MCM 的应用

MCM 可以应用在很多场合，不过在使用 MCM 时必须注意以下几点。

1）首先必须能确保 MCM 外形比单芯片封装更小。通常，实现理想特征尺寸的最有成效的方式就是将数字元器件集成到 MCM-L 中。如果需要最小的尺寸，可以采用 MCM-D 技术来实现。

2）另一种可选的方案是采用单冲模。不过单冲模可能会太大而无法生产，或者太大以至于效率太低。其设计可以根据用户进行定制或半定制。这种情况下，可以将冲模划分为很多更小的冲模，并利用 MCM 互连来实现大块冲模的性能，这也是一种不错的选择。

3）我们可以采用混合技术来控制单片 IC 的成本以及电气性能。例如，我们必须将互补金属氧化物半导体（CMOS）数字 IC 和 GaAs 微波单块集成电路（MMIC）或高带宽模拟 IC 连接在一起；否则，我们就必须采用大量的静态随机存取存储器（SRAM）和少量的逻辑电路来配合使用。这种情况下，MCM 就变得非常实用了。如果互连芯片导线很少，那么 MCM-L 可能是最好的选择。如果需要的布线密度很大，那么多层 MCM-C 或 MCM-D 可能是最好的选择。

4）如果采用单芯片，那么 IC 之间的焊接点或 IC 速度就会受到限制。例如，很多计算机设计就是得益于两个高速缓冲存储器之间很宽的总线（256 位或更大）。这种情况下，MCM 就允许在 IC 之间存在很多很短的连接。

5）如果采用单芯片封装，我们就不能确保能否实现理想的电气性能。例如，总线运行速率高于 100MHz 时。如果仿真过程表明很难保证总线的速度，那么就可以采用 MCM 了。另外，例如在混合信号设计中，其噪声就很难控制。一般在单芯片封装设计中，MCM 可以提供较高级别的噪声等级和互连速度。

6）传统的设计中包含很多焊接点，其可靠性对我们来说很重要。而 MCM 的设计包含的焊接点就少了很多，因此 MCM 也少了很多感应产生的影响，从而不合格的产品就很少（倒装晶片的焊球凸点比板级焊接点具有更高的可靠性）。

尽管 MCM 技术还可用于其他方面，但是到目前为止，上述这些还是最主要的应用。如果采用 MCM 技术，下一个问题就是确定 IC 如何放置到 MCM 上，下面给出了几点建议：

1）首先尽量使 MCM 的封装与单芯片的封装安装技术相互匹配，以提高制造的效率。

2）尽管 MCM 中芯片之间布线的成本不高，但是远离 MCM 的引脚是很昂贵的，因此，MCM 中应尽可能地多包含一些内部布线。另一方面，一个过于复杂而且包含很多元器件的 MCM，其效率也会大打折扣。

3）在混合信号的设计中，将产生噪声的数字元器件与敏感的模拟元器件分隔开是很有必要的，这种分隔可以通过将数字元器件放置在MCM中来实现。如果采用包含集成去耦电容的MCM-D，那么MCM产生的噪声将会很小，这样数字元器件和模拟元器件就可以很好地在一起协同工作。但是MCM的接地特性将和PCB的接地特性有很大区别。

简而言之，大多数MCM系统的设计要素主要取决于系统级的成本和性能的建模过程。虽然MCM技术的封装成本较高，但MCM在系统其他方面节省了不少成本，因此，MCM还是非常理想的选择。

13.5 MCM的设计要点

MCM设计中最重要的事情就是获取裸芯片。每个MCM系统中涉及到的第一个问题就是是否存在足够多的理想冲模，其测试等级是否合适，是否有补充货源以及合适的价格。解决这个问题只能依靠时间，因为许多芯片厂商时刻在观察他们的冲模销售情况，以瞄准市场的各个机会。另外，如果生产出来的冲模无法使用，要尽快确定可替代的芯片来源。

第二个要点就是测试和检验方案。MCM和PCB有很多不同的地方。首先，对芯片模块来说，由于芯片模块设计的成本很高，因此一次性通过测试和检验的希望就很大。如果采用MCM，完整的预制造设计检验就显得更为重要，因此在制造之前的逻辑和电气仿真过程上就必须下很大的功夫。

在设计过程中，如何在安装好的MCM中诊断出故障也是很重要的。在原型设计中，我们希望可以在重新设计之前就能对设计错误进行定位。在实际的产品模块中，如果故障模块需要维修（通常是替换一个冲模），就必须确定故障冲模的位置或者导线连接点/焊球凸点的位置。但是，在MCM上比PCB上更难查找出导线的故障点，因此，必须准确制定并实施故障隔离测试方案。测试方案必须能将故障与单芯片或者芯片间的互连部分相隔离。当采用具有边界扫描功能（见图13-11）的芯片时最好采用这种测试方案。通过边界扫描，测试矢量就可以连续扫描到每个芯片周围的寄存器。之后，MCM就可以运行一个时钟周期，并得到扫描输出的结果。扫描输出结果可以用来确定出现故障的芯片或连接部分的位置。如果边界扫描无效，那就需要对MCM进行维修了，然后还要制定一个芯片间故障检测的替代方案。

是否需要进行测试主要取决于成本和测试中断率（Sandborn和Moreno在1994年的著作以及Ng in Donane和Franzon在1992年的著作第4章中介绍了相关的信息）。通常，如果MCM主要由便宜、成熟的冲模构成，那么就没有必要进行维修，因为维修的成本已经超过了一个新模块的成本了。对于采用高成本、低

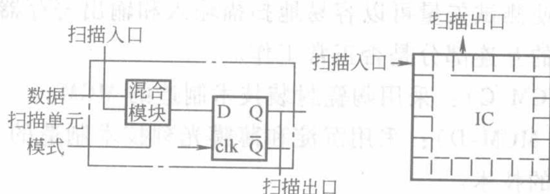


图 13-11 边界扫描的应用可以使 MCM
中故障的定位变得更容易

效率冲模的 MCM 来说，也是同样的道理，尤其是当确定故障就是由于那个冲模导致的之后，就更加没有维修的必要了。另一方面，故障诊断和维修通常是对于那些包含很多高价值冲模的模块来说的，因为当其中一个冲模出现问题时，不可能将所有的冲模都扔掉。

热学设计在 MCM 中也非常重要。一个 MCM 比功能相同的 PCB 具有更高的散热密度，因此也很有必要寻找一个更复杂的散热方案。如果 MCM 的散热功率超过 1W，那么就必须检查是否需要吸热设备或散热片了。

有时，在 MCM 中热量集中也可以成为设计师利用的一个优势。相比单芯片封装中很多块散热片的情况，如果 MCM 采用一个很大的散热片，那么就可以进一步节省成本了。

通常，MCM 中的电气设计比 PCB 中更容易，因为芯片间的连接更少，而且寄生电感和电容更小。如果采用 MCM-D 技术，可以将电源和接地点布置得很近，以便产生一个完美的去耦电容。但是，MCM 中的电气设计必须仔细进行，因为例如 300 MHz MCM 的设计复杂度与 75 MHz PCB 的相当。

MCM 通常用于混合信号的设计（混合了模拟和数字信号）。混合信号 MCM 电气设计的要点类似于混合信号 PCB 的设计要点，而且相关的辅助设计工具也很有限。目前，MCM 设计已经被定性了。还有一件很重要而且只与混合信号 MCM 的设计有关的事实，那就是 MCM 的电源和接地与 PCB 中的电源和接地是通过寄生效应相互隔离的。很多设计师都认为 MCM 和 PCB 的参考电压只有在原型中查找噪声问题时才会是一致的。

更多关于 MCM 设计技术的信息，读者可以参考 Donane 和 Franzon 在 1992 年、Tummula 和 Rymaszewski 在 1989 年以及 Messner 等人出版的著作。

名词解释

边界扫描：芯片在安装到 PCB 或 MCM 之后，用来测试芯片的技术。芯片会

进行特定设计，以便测试矢量可以容易地扫描输入和输出寄存器。这些矢量用来确定芯片、芯片间的互连部分是否正常工作。

陶瓷 MCM（MCM-C）：采用陶瓷封装技术制造的 MCM。

薄膜式 MCM（MCM-D）：采用沉淀和薄膜光刻技术制造的 MCM，类似于集成电路制造中采用的技术。

倒装晶片焊球凸点：芯片焊接技术，在芯片的焊接点和表面上有一个焊球。然后，芯片就会和 MCM 或 PCB 配合使用，焊料就会回流形成一个焊接点。该技术允许使用区域附属技术。

知名优质冲模（KGD）：经过高级别的测试的裸硅片（冲模）。

碾压 MCM（MCM-L）：采用先进的 PCB 制造工艺制造的 MCM。

多片组件（MCM）：一个单独的、包含很多芯片的、预先组装好的元件。

印制电路板（PCB）：大多数电气设备中都包含的传统电路板。

带状自动粘结（TAB）：一种制造工艺，在该工艺中，引脚被冲压成金属带状，芯片就是被焊接在引脚的末端内侧；然后芯片和引脚焊接框架被装配到 MCM 或 PCB 上。

导线连接：芯片从其每个焊接点引出一条引线来焊接到 MCM 或 PCB 上相应的焊接点。

感谢

在此，作者要感谢 Andrew Stanaski 的校对和图 13-11 的信息以及相关的注释。同时，作者还要感谢 Jan Vardaman 和 Daryl A. Doane，他们提供了非常宝贵的学术经验，其中很多都在本章中用到了。

参考文献

[1] Dehkordi, P., Ramamurthi, K., Bouldin, D., Davidson, H., and Sandborn, P. 1995. Impact of packaging technology on system partitioning: A case study. In 1995 *IEEE MCM Conference*, pp. 144-151.

[2] Doane, D. A. and Franzon, P. D., eds. 1992. *Multichip Module Technologies and Alternatives: The Basics*. Van Nostrand Reinhold, New York.

[3] Franzon, P. D., Stanaski, A., Tekmen, Y., and Banerjia. S. 1996. System design optimization for MCM. *Trans. CPMT*.

[4] Messner, G., Turlik, I., Balde, J. W., and Garrou, P, E., eds. 1992. *Thin Film Multichip Module*. ISHM.

[5] Sandborn, P. A. and Moreno., H. 1994. *Conceptual Design of Multichip Modules and Systems*. Kluwer, Norwell, MA.

[6] Tummula, R. R. and Rymaszewski, E. J., eds. 1989. *Microelectronics Packaging Handbook*. Van Nostrand Reinhold, Princeton, NJ.

备注

关于多芯片模块设计技术，读者还可参考以下资料。

其中，大部分书籍位于参考文献中；而刊物主要包括：《*IEEE Transactions on Components*》、《*Packaging and Manufacturing Technology*（A 和 B）》和《*Advancing Microelectronics*》，由 ISHM 出版。

其中最重要的商业杂志是：《*Advanced Packaging*》，对会员免费。

两个主要的技术会议为：（1）IEEE 多片组件会议（Multichip Module Conference，MCMC）；（2）ISHM/IEEE 多片组件会议。

第 14 章　集成电路的测试

Wayne Needham

14.1　引言

集成电路测试的目的是为了将合格器件与不合格元器件区分开，并确定哪些是合格器件。测试显示为不合格的不合格器件，则成为生产中的损耗，但是当我们减少生产过程中不合格产品数量时，这就意味着出现降低成本的机会。如果合格器件被检测为不合格器件，那就会出现浪费的现象，从而会直接影响到成本和利润。最后，如果不合格器件被检测为合格器件，那就会产生质量问题，通常用“每百万个器件中存在缺陷的数量（Defects Per Million，DPM）”来描述。

不过，测试环境和工作环境是不完全相同的。测试过程很可能拒绝合格产品而将有缺陷的产品认定为合格产品，这就是我们当前测试方法中的一个基本问题；这个问题我们必须彻底理解并分析它是如何产生的，以便我们更好地进行测试。图 14-1 给出了这种交叉测试错误的情形。

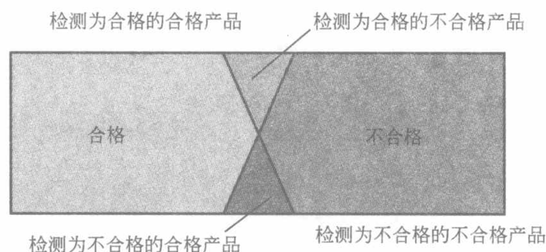


图 14-1　交叉测试错误

不合格的器件只有通过一系列的测试（例如基于电压或电流）才能从合格产品中筛选出来。具体的测试过程和实际应用系统本章将不做介绍。

14.2　缺陷类型

集成电路中的缺陷主要包含 3 种类型，具体描述如下：

1）信号线路和电源上出现的开路或异常高阻部分。这种缺陷会增加连接点

之间的串联电阻。这种多余的电阻会减慢该信号线路上寄生电容充放电的速度，并导致时延。这种缺陷还会使电压衰减至晶体管电路的水平，而晶体管电路也会产生时延。

2）集成电路中各层之间出现的短路、桥接或者高阻连接。这种缺陷在正常的绝缘区域形成了一条高阻路径。如果其 R 值足够小的话，该路径就无法传输信号 1 或者 0 了。

3）参数发生的变化（例如由于污染、微粒、注入或者工艺参数变化时导致的开启电压和饱和电流值偏离）。这种缺陷通常会导致电路速度或驱动电压发生变化。

表 14-1 给出了各种缺陷的分类以及这些缺陷在集成电路中产生的原因，表中同时还描述了缺陷的电气特性。

表 14-1 典型的 IC 缺陷和故障模式

缺陷类型	产生原因	电气特性
多余的金属或互连	微粒	导线之间出现电阻连接
受限制的几何尺寸	本身缺少或由于蚀刻导致	高于正常的阻值
生产过程中的外来材料	污染	变化的电阻值或短路

尽管集成电路在生产过程中出现一些缺陷是很常见的事，但是这些缺陷很难通过测试方法来量化。因此，工业上将这些缺陷称为“故障”，故障检测过程在最大程度上模仿出缺陷在测试环境中的特性。

传统故障及故障模型

应用最广泛的故障模型是“单锁定故障模型”，该模型具有很好的支持工具，可用于各种仿真环境。这种模型描述了信号线路至电源线之间的一种短路缺陷，这种缺陷导致信号始终是逻辑 1 或 0。因此，其处理过程的逻辑结果就无法传送至输出端。图 14-2 给出了单锁定故障模型的逻辑电路图。尽管这种故障模型很常见，但是该模型早在 30 年前就已经开发出来了，而且并不比当前更加先进的工艺落后很多。当前超大规模集成电路（VLSI）和亚微工艺通常在信号线路上也存在一些桥接缺陷，这种桥接缺陷在单锁定故障模型中不能很好地反映出来。

其他推荐的有效故障模型概括如下。

1）开路故障模型：该模型适用于开路。不过，在 CMOS 中，即使缺少部分接触点和晶体管，但 CMOS 在切换过程中性能仍然表现正常。

2）泄漏故障模型：该模型适用于晶体管中有泄漏的情况，采用 I_{DDQ}[⊖]测试技术（本章稍后将会讨论）。

⊖ I_{DDQ}是 CMOS 电路中商电流的 I_{IEEE} 标志。

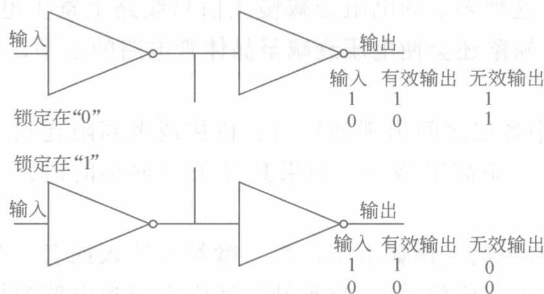


图 14-2 单锁定故障模型

3）定时故障模型（时延、门电路和晶体管缺陷）：该模型适用于电路中 AC 特性的变化。电路中的单门电路信号、线路信号或者时钟信号比实际需要的或仿真的信号慢，这种慢信号会在高速运行的集成电路中产生问题。

4）假锁定 I_{DDQ} 故障模型：该模型适用于那些电平看上去像固定在逻辑 1 或 0 的节点。

5）表决故障和桥接模型：该模型适用于相邻的线路桥接和逻辑结果未知的情况，或者具有很强驱动能力的表决电路支配了逻辑电平的情况，同时还适用于存储器电路。

6）邻近模式故障模型：该模型适用于一个存储器单元的运行会影响到邻近存储器单元的情况（通常局限于 4 个相邻的存储单元）。

7）耦合故障模型：该模型适用于满行或满列将信号耦合进存储单元或相邻行或列中的情况。

8）保留故障模型：该模型类似于逻辑开路故障模型（如前所述），该故障模型中的故障不会将数据保留很长时间。

故障模型还包括其他很多类型。

有一点很重要，即锁定模型不一定能捕捉到制造过程中产生的故障或缺陷，这一点目前正受到计算机辅助设计（CAD）业界的关注。目前，有很多解决方案都有效，或者至少可以给出故障模型的各个要点。目前也有很多有效的工具能够支持时延故障模型。桥接故障模型的流程已经通过了大学研究机构的论证，在未来几年内将可以在 CAD 提供商的工具中应用了。不久之后，除了锁定模型，我们将可以通过其他模型来进行测试了。

14.3 测试的概念

集成电路的主要测试方法就是控制和观察节点，这样可以验证逻辑工作情

况，以确保电路没有故障。这个过程通常是通过产生输入激励模式并比较输出信号和有效状态来实现的。图 14-3 给出了对集成电路中的小部分电路进行测试的经典示例，该测试中输入为激励模式组，输出为理想状态组。注意到该测试电路是逻辑电路（而不是实际布线的电路板），不是独立的；而且只适用于锁定模型。

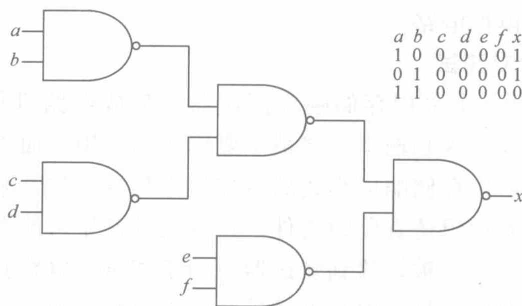


图 14-3 逻辑图和测试组

1. 测试类型

集成电路主要包含 3 种基本的测试类型，分别为参数测试、电压测试和电流测试。每一种测试都在测试过程中占有重要地位。

参数测试的目的是用来确保集成电路中晶体管的正常工作（电压和电流）。参数测试包括定时、输入输出电压和电流测试。通常，这些测试主要针对电路的性能。所有的测试中，测试结果都和测试方法有关。典型的测试值如：$T_x < 334$ ns，$I_{OH} < 4$mA，$V_{OI} < 0.4$V 等等。

下一个主要测试类型为电源电流测试。电流测试包括有源切换电流或电源测试、电源下降测试以及等待或稳定静态电流测试。其中，最后一个测试通常称为“I_{DDQ}”，稍后将对其进行讨论。

集成电路中最常见的测试方法是基于电压的测试。逻辑操作中基于电压的测试包含了各种情况，例如在集成电路的输入端设置 1，并确保 IC 的输出端在正确的时间输出合适的 1 或 0。在大多数情况下，逻辑值会在高和低工作电压之间重复。图 14-3 就是一个基于电压测试的例子。

2. 测试方法

为了进行基于电压的测试或者其他类型的测试，必须对电路进行初始化，并对内部节点进行控制和观察。下面列出了 4 种基本的测试方法：

1）外部存储响应测试：这种测试方法是目前集成电路最常见的测试形式，主要依赖于大型 IC 自动测试设备（Automated Test Equipment，ATE）。

2）内置自行测试（BuiltIn-Self-Test，BIST）：这种测试方法在集成电路上定义并

构建测试电路。这种方法提供了激励模式，并对集成电路的逻辑输出值进行观察。

3）扫描测试：部分或全部时序单元被转换成移位寄存器，用于控制和观察。

4）参数测试：直接测量集成电路的电路参数。这种测试包含了 I_{DDQ} 测试，I_{DDQ} 测试善于检测某些类型的缺陷。为了进行 I_{DDQ} 测试，必须终止 CMOS 集成电路的时钟，然后再对静态电流进行测量。所有有源电路必须被置于一个低电流的状态，包括所有的模拟电路。

3. 外部存储响应测试

图 14-4 给出了一个采用存储响应测试器对集成电路进行测试的典型示例。注意到，集成电路的输入和输出模式被主要输入端采用，而输出端将会与知名有效响应进行比较。产生存储响应模式的过程通常是通过仿真来实现的。一般，这些模式是电路的原始逻辑仿真曲线文件。这些文件开始可能在正常工作逻辑电路的仿真过程中使用，然后被转移到测试器中用于调试、电路验证和测试。这些模式可能占用很大的存储容量，但是测试模式可以很轻易地检测出集成电路中的结构性或功能性缺陷。

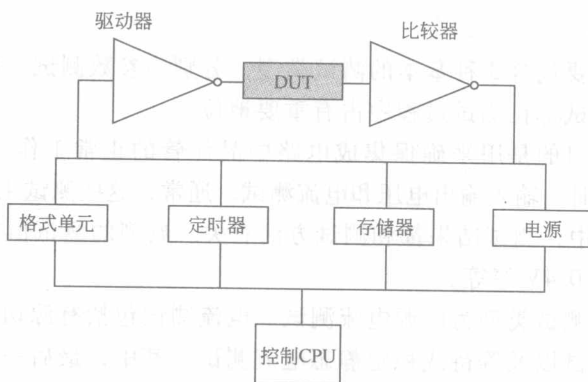


图 14-4 存储响应测试器和 DUT

为了对待测器件（Device Under Test，DUT）进行测试，存储响应功能测试器必须依赖以上介绍的各种逻辑仿真模式。这些模式中包含了输入和输出状态，而且必须说明未知状态。功能测试的模式集通常可以是曲线文件或者变化文件，然后这些文件就会被输入到测试设备中。曲线文件或者变化文件将 DUT 的逻辑状态捕捉下来作为其整体的逻辑输出变化。例如，这些模式可以说明待测设备的逻辑操作情况。同样，这些模式也可用于集成电路的设计和运行调试。不过，这些模式不适合故障隔离或器件的分析。

4. BIST

最常见的 BIST 实现方式包括采用线性反馈移位寄存器（Linear Feedback

Shift Register，LFSR）作为输入端。LFSR 的结构来自移位寄存器，其最低有效位被反馈到具有专用 OR 门的选定电路。这种反馈机制产生了一个多项式，该多项式逐次除以商并将余数相加。只有某些多项式才能产生伪随机模式，因此多项式的选择必须仔细考虑。进行初始化和时钟设置时，LFSR 源会产生一个多项式模式，该模式可以作为激励模式用于集成电路中的逻辑模块。

常见的输出压缩测试方法采用了多输入移位寄存器（Multiple Input Shift Register，MISR）。MISR 是一种 LFSR，其逻辑时钟模块的输出端连接到具有专用 OR 门的 LFSR 电路。如果运行顺畅，逻辑模块将会产生一个正确的单信号（见图 14-5）。如果逻辑模块存在缺陷，那么在输出 MISR 端将会捕捉到缺陷产生的逻辑误差。MISR 始终采用选定电路的反馈值来除以输出值，这样误差就可以一直保留在 MISR 中，直到最后的结果在测试顺序的末端被读取。由于逻辑状态被压缩在 MISR 中，因此如果在输出端误差不断重复或者在模块中存在多个误差，那么最后产生误差的原因可能会是掩模缺陷。尽管可能存在很多误差，但即使逻辑模块中也包含误差或存在缺陷，输出值中仍然可能包含正确的输出值，这种现象称为“混淆现象”。例如，如图 14-5 所示的输出寄存器为 20 位长，2^{20} 约为一百万，这就是说出现混淆缺陷的几率非常小，集成电路的输出值为正常状态的机率非常大。

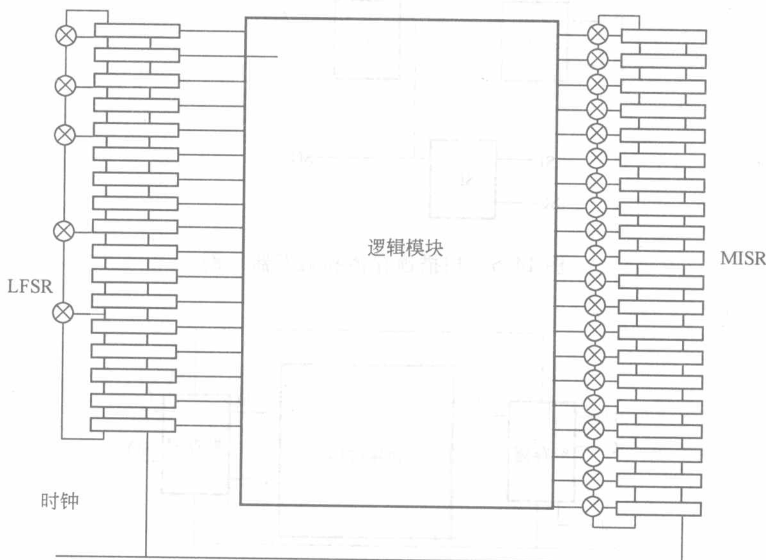


图 14-5 BIST 输入和输出端的实现方式示例

我们必须注意，图 14-5 中的例子只适合组合逻辑模块。时序逻辑在初始化后会变得更加复杂，因此在 BIST 设计和模式设计中，必须考虑无效状态、状态

机、全局反馈和许多其他因素。

5. 扫描测试

扫描是一种将存储单元（锁存器和触发器）变成双模式单元的技术。例如，图 14-6 给出了集成电路中一个常见的触发器例子。在图 14-6 中，触发器被转换成一个扫描寄存器单元。在常规操作中，D 输入端和系统时钟控制着触发器的输出值。在扫描操作中，扫描时钟用来控制数据在移位寄存器中的进出，数据被移位进出主锁存器和扫描锁存器。次锁存器由系统时钟控制，以便在本级扫描锁存器和下一级扫描锁存器之间的逻辑电路中产生一个所需的激励信号。图 14-7 给出了锁存器和锁存器之间的逻辑电路的示例。这种方法有效地将复杂的顺序测试问题简化成了一个简单的组合问题。不过，在扫描过程中还存在一个问题，即电路占用总面积是目前所有测试设计（Design For Test，DFT）方法中最大的。这个面积的增加是由于必须增加每个存储单元中晶体管的数量来实现扫描中的加法功能而造成的。

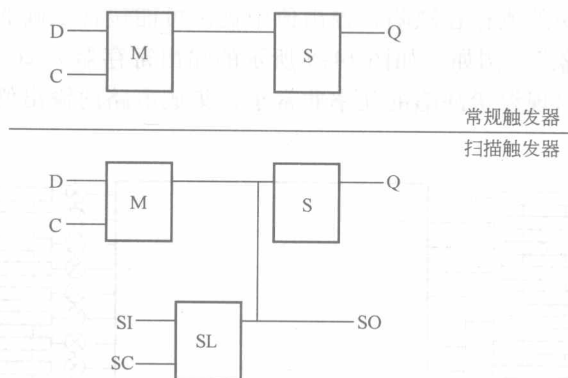


图 14-6 扫描锁存器和触发器示例

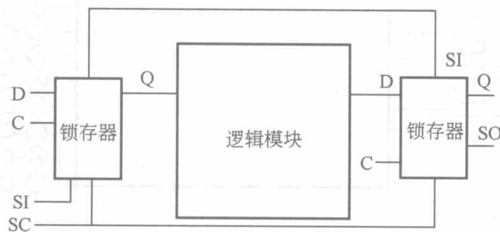


图 14-7 逻辑电路中扫描锁存器的使用

注意到，扫描锁存器中包含了额外的控制和观察单元。这些额外的单元包括扫描时钟（Scan-Clock，SCK）、扫描输入（Scan-In，SI）和扫描输出（Scan-

Out，SO）。在图 14-7 中，扫描时钟用来驱动所有的扫描锁存器，并控制着扫描寄存器的移位寄存器功能，并允许为扫描链上的每个锁存器设置 1 或 0。每次扫描时钟在应用时，扫描输入线上将会被置 1 或 0，该值然后会被转移至输出端和下一个锁存器的扫描输入端。当整个扫描链被初始化时，这个转移就停止了。在顺序扫描的末端，系统时钟会被触发，并从 Q 输出端中锁存器的扫描部分向外发送数据。扫描完成后，系统将会加载并转移出存储的逻辑操作值，传送给比较器。如果逻辑操作是正确的，那么扫描值在输出顺序中将会是正确的。

6. I_{DDQ}测试

I_{DDQ}测试可能是最简单的测试概念。图 14-8 给出了一个具有潜在缺陷的典型双反相器结构，该图还说明了 I_{DDQ}的特性。图中上半部分的电路是无缺陷的，而下半部分的电路具有泄漏缺陷，即 R 所示部分。I_{DDQ}的特性随时钟周期变化而变化。注意观察时钟周期 A 中有缺陷的电路其上升电流的变化，这个上升电流就称为“I_{DDQ}”。I_{DDQ}非常适合用来检测大多数桥接缺陷和其他类型的缺陷。每次测试中，I_{DDQ}测试控制并观察着约 IC 中约半数的晶体管。因此，I_{DDQ}测试可以同时检测几种缺陷，如可以有效检测桥接缺陷和部分短路缺陷。I_{DDQ}测试中还必须包含几个设计，例如当时钟停止时，所有的 DC 路径、上拉电路、总线驱动器以及竞争路径必须设计成零电流。如果只剩下一个单独的晶体管，就无法进行 I_{DDQ}测试。

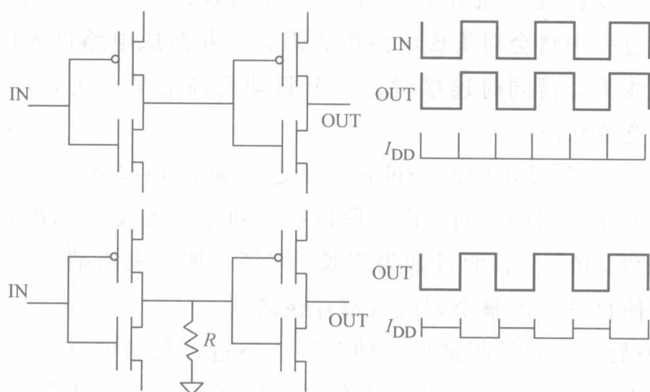


图 14-8 I_{DDQ}测试

注意到，在有缺陷的电路中，如果输出电压足够用来驱动下一个输入电路，那么 V_{out}和 V_{in}还是可以接受的。此处没有给出的是衰减电压对定时的影响，该衰减电压会在下一级电路中产生一个时延。

14.4 测试中的权衡

为了确定最好的测试方法，我们必须掌握缺陷类型、集成电路的逻辑功能和测试要求。表 14-2 给出了测试过程中权衡问题的示例。表中每行对应一个如前所述的测试类型：扫描、BIST、存储响应和 I_{DDQ}；表中每列描述了此处介绍的测试类型的属性。

表 14-2 测试的属性

测试类型	复杂度	无效状态	混淆现象	在硅片上的面积	测试模式形成时间
扫描测试	简单	受约束的	NA	最高	最短
BIST	与锁存器有关	大问题	有	中等	中等
存储响应测试	非常复杂	NA	NA	一点	最长
I_{DDQ}测试	低复杂	没有	没有	没有	简单

表 14-2 中，第一列是测试的类型，第二列是测试模式的复杂度，即集成电路中节点激励和控制必要模式的产生难度。第三列是无效状态。通常，集成电路中的逻辑模块（如双向总线驱动器和信号）必须设置一致，例如，译码器、总线驱动器和互斥线路驱动器电路。如果设计错误，测试方法（如扫描测试和 BIST）在测试过程中将会对集成电路造成损害，并造成电路进入内部竞争状态。例如，两个总线驱动器同时连接时，一个驱动至高电平，而另一个驱动至低电平，就会产生竞争状态。

混淆现象是 BIST 技术中的一种常见问题，如表 14-2 中第四列所示。例如，假设存储器排列在行和列之间，而且假设有一列完全无效。如果在各列之间存在多种故障（假设 256 个），而且如果字长为 256，那么输出端也可能产生一个正确的响应，即使存储器的整个列或行都存在缺陷。

表 14-2 中最后一列是测试形成的时间。不过，我们可以看出占用硅片面积最大的测试技术，同时也是对无效状态控制最好和形成测试模式时间最短的技术；而占用硅片面积最小的技术具有最复杂的测试形成时间问题，即需要最长的测试模式形成时间。这就是复杂的权衡问题。

在选择 DFT 和测试方法之前，对即将制造的集成电路的尺寸进行预测是很重要的。设计过程是临时的，而且可以被平均分配到整个集成电路的制造过程中去。集成电路的设计可以是很简单的过程，也可以是一个很复杂的过程。

测试技术的选择标准必须考虑集成电路的尺寸、预期性质和预期成本。出货和上市销售的时间是关键因素，因为某些大规模集成电路的测试技术会耗费很长的时间。

名词解释

自动测试设备（ATE）：计算机控制的、用于测试集成电路的设备。

内置自行测试（BIST）：一种设计技术，包含输入激励产生电路和输出响应检测电路。

缺陷：电路结构中的误差、微粒或污染。缺陷可能会导致错误的电路操作，也可能不会。

测试方法的设计（DFT）：该名词包含了所有用来提高内部节点控制和观察的技术。

待测器件（DUT）：等待 ATE 测试的器件。

故障：电路运行的错误。通常故障电路是由于缺陷或电路的损坏导致的结果。

线性反馈移位寄存器（LFSR）：构成寄存器的一种方法，其反馈用来产生伪随机序列。

多输入移位寄存器（MISR）：通常，一个 LFSR 的输入端由前一级的带有输入数据的专用 OR 等电路构成。这种方法将电路的响应结果缩简成一个多项式。

扫描：一种设计技术，在该设计技术中时序单元（锁存器或触发器）被串连成一种模式，该模式允许数据从锁存器中移进移出。通常来说，扫描过程允许对锁存器之间的逻辑电路进行简单访问，这样可以大大降低测试模式形成时间并简化测试工作。

参考文献

[1] Abramovici, M. et al. 1990. *Digital Systems Testing and Testable Design*. IEEE Press, New York.

[2] Needham, W. M. 1991. *Designer's Guide to Testable Devices*. Van Nostrand Reinhold, New York.

[3] van de Goor, A. J. 1991. *Testing Semiconductor Memories, Theory and Practice*. John Wiley&Sons. New York.

备注

读者还可以参考以下会议的相关资料。

[1] IEEE 国际测试会议（International Test Conference，ITC）：该会议聚集了全球最多的测试专家、测试提供商和研究人员。

[2] IEEE 设计自动化会议（Design Automation Conference，DAC）：该会议在全球各地举办。

[3] IEEE VLSI 测试座谈会。

各种类型测试设备的测试供应商还可以提供 ATE 的操作、编程和维护培训。

CAD 提供商提供了各种各样的仿真工具，用于测试和调试。

第 15 章 半导体故障模式[㊀]

Victor Meeldijk

15.1 分立半导体故障模式

半导体器件在存储应用（或待用状态）中存在的故障，是由于潜在的制造缺陷在器件筛选测试过程中没有被发现而导致的结果。对于分立半导体来说（例如晶体管），很大一部分的故障是由于冲模和导线连接的缺陷与污染导致的结果。弹簧二极管的常见故障模式就是接触材料失去了压缩强度或者滑落芯片而最终导致的开路。

故障类型主要可以划分为以下 3 种：

1）与环境无关的故障（氧化缺陷、扩散缺陷）；

2）与环境相关的故障（导线连接或镀金缺陷，如图 15-1 和图 15-2）；

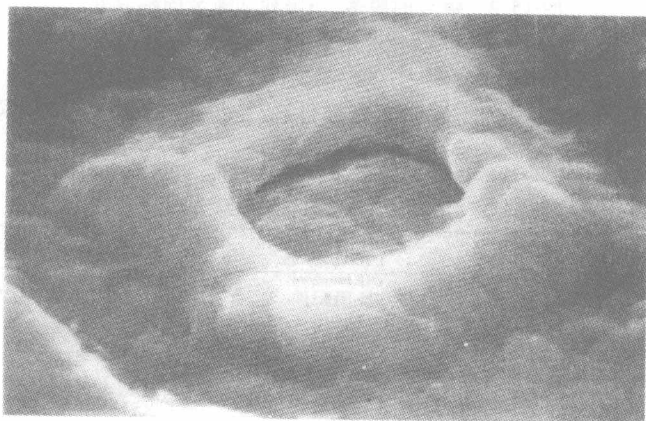


图 15-1 过度的侵蚀或镀金

3）与环境和时间都相关的故障（金属发生移动、侵蚀、金属互化物，例如由于使用不同类型的金属而导致的故障）。

㊀ 本章的部分内容是改编自：Meeldijlk，V. 1995. *Electronic Components*：*Selection and Application Guidelines*，第 10 章和第 11 章。Wiley Interscience，New York 出版，引用经过允许。

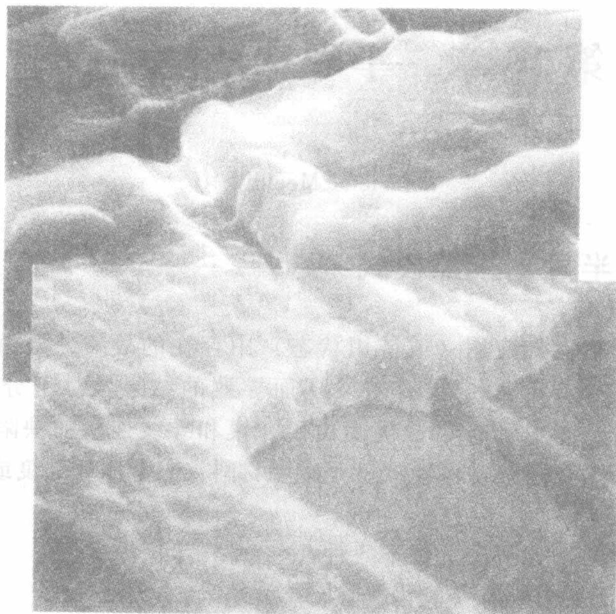


图 15-2 镀金的脱落/凹陷处和镀金的稀薄处
（第二张照片中是可接受的镀金效果）

表 15-1 和 15.6 节中讨论了在元器件安装之前可能产生故障的各种筛选测试。

表 15-1 用来检测缺陷的各种筛选测试方法

测试方法	缺陷处理	故障模式
电气测试： DC 测试、静态/动态测试、功能性测试、Schmoo 结构测试	备用配件 蚀刻 导线连接 对错误进行标记	错误的电气特性、开路、短路、退化的结特性、间歇性故障 与期望的工作特性相差甚远，或者根本不同
热循环： MIL-STD-883（测试标准）方法 1010；日本工业标准（JIS）C 7022 A-4；国际电工技术委员会（IEC）PUB 68；测试 Na，Nb （温度/湿度循环 MIL-STD-883 方法 1004；JIS C 7022 A-5；IEC PUB 68；测试 Z/AD）	备用配件 钝化（见图 15-7） 蚀刻、扩散 镀金 DIE 分隔 DIE 连接 导线连接、密封 老化加热分解	潜在的短路和开路、低击穿电压、增加的泄漏、由于故障元器件导致的性能减退、近似短路、近似开路、高阻性内部连接

（续）

测试方法	缺陷处理	故障模式
高温存储： MIL-STD-883，方法 1008；JIS C 7022，B-3 高温/高湿偏置测试和存储测试，JIS C 7022，B-5；IEC PUB 68；测试 C （高温测试标准：MIL-STD-750 方法 2031；JIS C 7022 A-1；IEC PUB 68；测试 Tb） （高温操作标准：MIL-STD-883 方法 1005；JIS C 7022 B-4；IEC PUB 68；测试 A）	蚀刻 备用配件 扩散、镀金 引线连接、密封 徐变、侵蚀 电迁移、表面电荷扩展	由于结的迁移导致器件性能出现退化（低温焊接特性、短路、潜在的短路、开路、高阻）
反向偏置电压：	备用配件 密封	结性能退化、电路性能退化
工作寿命测试：	备用配件 钝化 扩散、镀金 冲模粘结、密封 树枝状结晶生长	结性能退化、短路、泄漏、低绝缘性、由于加热或者缺陷或不稳定的内部元件导致的性能退化、开路、近似开路、近似短路、高阻性内部连接
机械测试： 可变频率振荡标准：MIL-STD-883 方法 2007；JIS C 7022 A-10；IEC PUB 68；测试 Fc	冲模分离 冲模粘结、导线连接 冲模分裂、引线定形 不规则封装、不规则徐变	由于过热导致的性能退化、开路或开路隐患、短路或间歇性短路
可焊性： MIL-STD-883；方法 2003；JIS C 7022 A-2 盐性雾气（气体） MIL-STD-883 方法 1009；JIS C 7022 A-12；IEC PUB 68；测试 Ka	除铅/电镀	间歇性故障、开路或元器件与电路板之间的高阻性焊接 测试在盐性雾气侵蚀中的阻抗特性
热冲击： MIL-STD-883 方法 1011；JIS C 7022 A-3；IEC PUB 68；测试 Nc	冲模分离 冲模粘结	由于过热导致的性能退化、开路或潜在的开路、短路或间歇性短路
X 射线：	冲模粘结、导线连接 密封 引线框不规则定型	由于过热导致的性能退化、间歇性故障、开路或短路、由于封装中稀松的导电微粒导致的间歇性短路

（续）

测试方法	缺陷处理	故障模式
加速： 均匀加速，MIL-STD-883 方法 2001；JIS C 7022 A-9；IEC PUB 68；测试 Ga	冲模粘结 密封	由于过热导致的性能退化、易断线路的断开或间歇性断开、由于封装中稀松的导电微粒导致的间歇性短路
高电压测试：	钝化 密封	开路或短路、增加的元器件泄漏、低绝缘击穿
ESD 测试： MIL-STD-883 方法 3015；日本电气工业协会（EIAJ）IC 121：20	ESD 设计的灵敏度	元器件性能的退化或元器件故障
振荡/PIND 测试：	冲模粘结 导线连接	短路或由于封装中稀松的导电微粒导致的间歇性短路、由于过热导致的性能退化、开路或性能故障
引线老化： 引线完整性；JIS C 7022 A-11；IEC PUB 68；测试 U	密封	开路、内部引线疏松
泄漏测试： 真空密封测试：MIL-STD-883 方法 1014；JIS C 7022；A-6；IEC PUB 68；测试 Q 高压容器测试（PCT），用来评估瞬间的潮湿阻抗；EIAJ IC-121：18	密封	性能退化、由于化学腐蚀或潮湿导致的短路或开路、间歇性故障
封装前外观：	备用配件 钝化、制模 蚀刻、电镀 冲模分离 冲模粘结、导线连接	增加的泄漏、低绝缘崩溃、开路、短路、间歇性故障、潜在短路或开路、高内部阻抗、间歇性短路
外观：	密封	开路、封装裂缝、封装密封问题

分立半导体的常见故障模式包括齐纳二极管的故障比例：50% 的故障为短路，50% 的故障为开路；结型二极管的故障比例：60% ~70% 的故障为反向；20% ~25% 的故障为开路，10% ~15% 的故障为短路；硅控整流器（SCR）的故障比例：2% 的故障为开路，98% 的故障为短路；晶体管的故障比例：20% 的故障为泄漏，20% 的故障为低增益，30% 的故障为开路，30% 的故障为短路（见图 15-3 所示）。

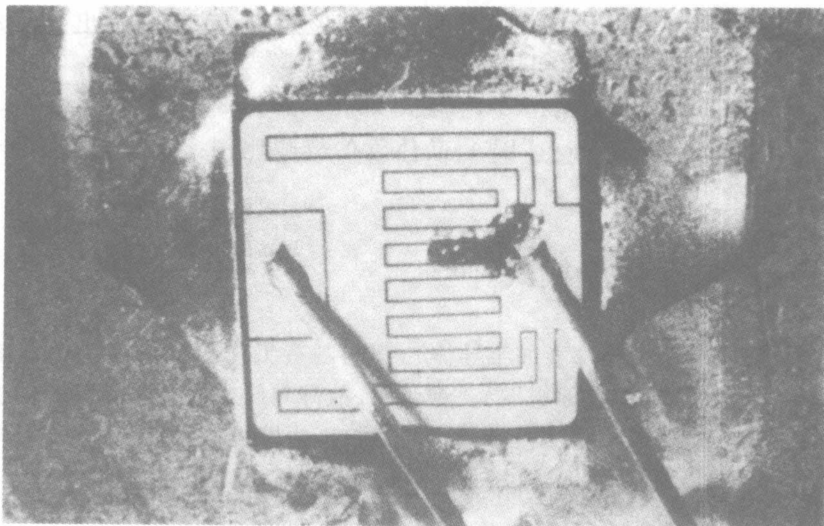


图 15-3 无效晶体管

在罗马和纽约的可靠性分析中心（Reliability Analysis Center，RAC，1991）关于故障数据的描述中，包含以下详细分类的故障模式：

1）常用二极管的故障比例：49%短路、36%开路、15%参数发生变化；

2）二极管整流器的故障比例：51%短路、29%开路、20%参数发生变化；

3）小信号二极管的故障比例：18%短路、24%开路、58%参数发生变化；

4）微波二极管的故障比例：50%开路、23%参数发生变化（偏离）、10%短路、17%间歇性故障；

5）齐纳二极管基准的故障比例：68%参数发生变化（偏离）、17%开路、13%短路、2%间歇性故障；

6）齐纳二极管整流器的故障比例：45%开路、35%参数发生变化（偏离）、20%短路；

7）光敏 LED 的故障比例：70%短路、30%开路；

8）光电传感器的故障比例：50%短路、50%开路；

9）硅可控整流器的故障比例：45%完全无效、40%短路、10%开路；

10）触发晶体管的故障比例：90%完全无效、10%部分无效。

可靠性分析中心（RAC）关于晶体管故障模式更详细的描述如下：

1）双极性晶体管的故障比例：73%短路、27%开路；

（注意观察这些新的数据与美国陆军装备司令部在 1976 年出版的数据（AMCP-706-196）的比较；在 1976 年的数据中，具有 59%的故障是基于泄漏电流

的，37%的故障是基于发射极崩溃的（Bvceo），剩下的4%是开路故障。）

2）场效应晶体管（FET）的故障比例：51%短路、22%低输出、17%参数发生变化、5%开路、5%高输出；

3）GaAs FET的故障比例：61%短路、26%开路、13%参数发生变化；

4）射频（Radio Freqency，RF）晶体管的故障比例：50%参数发生变化、40%短路、10%开路。

15.2 集成电路故障模式

大多数集成电路的故障（如半导体）都与制造工艺的缺陷有关（如图15-4所示的就是可接受的IC内部视图）。典型IC故障可详细分为：40.7%的引线互连、23.4%误用、4.2%制模或蚀刻缺陷、3.3%冲模机械损坏、1.4%裂开的冲模、0.9%冲模镀层被侵蚀、0.9%冲模污染和24.8%其他因素。

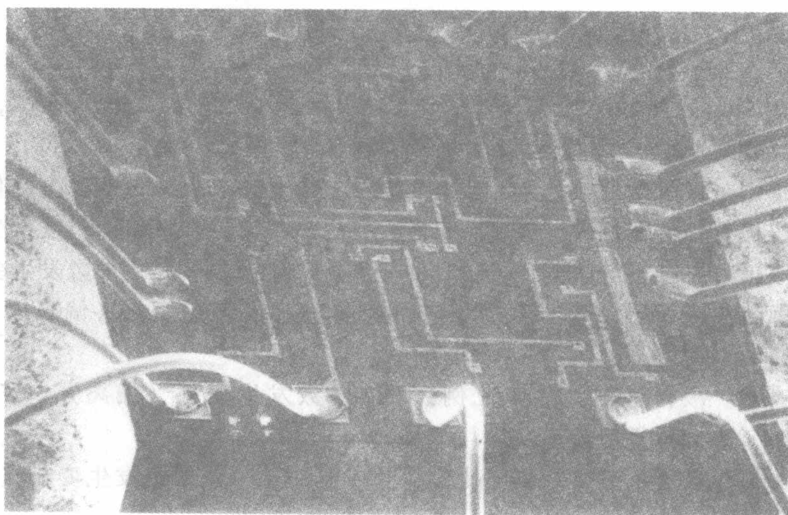


图15-4 可接受IC的内部视图，可以观察到粘结点和电镀层

这些故障因素会导致下列故障模式的产生：

1）数字器件的故障比例：40%锁定高电平、40%锁定低电平、20%失去逻辑功能；

2）线性器件的故障比例：10%～20%偏离、10%输出过高或过低、70%～80%没有输出；

可靠性分析中心在1991年公布的关于故障数据的描述中，故障模式比例如

下所示：

1）数字二极管的故障比例：28% 输出锁定高电平、28% 输出锁定低电平、22% 输入开路、22% 输出开路；

2）数字 MOS 的故障比例：8% 输出锁定高电平、9% 输出锁定低电平、36% 输入开路、36% 输出开路、12% 电源开路；

3）数字可编程逻辑阵列（PAL）的故障比例：80% 真值表无效、20% 短路；

4）接口 IC 的故障比例：58% 输出锁定低电平、16% 输入开路、16% 输出开路、10% 电源开路；

5）线性 IC 的故障比例：50% 性能退化/错误输出、41% 没有输出、3% 短路、2% 开路、2% 偏离；

6）线性运算放大器的故障比例：69% 性能退化（不稳定、截短的输出、偏离等等）、13% 间歇性故障、10% 短路、6% 瞬间过载、3% 没有输出；

7）双极性存储器的故障比例：79% 数据传输缓慢、21% 数据丢失；

8）MOS 存储器的故障比例：34% 数据丢失、26% 短路、23% 开路、17% 数据传输缓慢；

9）数字存储器的故障比例：30% 单比特错误、25% 列错误、25 行错误、10% 行和列错误、10% 完全无效；

10）数字 RAM 的故障比例：25% 在低温时无法工作、17% 参数发生变化、16% 短路、13% 开路、13% 数据错误、7% 污染（真空密封单元的腔内有微粒）；

11）紫外线可擦除可编程只读存储器（UVEPROM）的故障比例：94% 开路（不可编程）、6% 不能擦除；

12）混合器件的故障比例：51% 开路（由于电阻/电容开路、内部焊接点腐蚀和电迁移导致）、26% 性能退化/输出错误（变形或响应时间输出缓慢）、17% 短路（静电放电（ESD）过载、破裂、引线短接）、6% 没有输出。

这些分析数据来自可靠性分析中心（RAC，1991），可供读者在实际应用中作参考。大多数供应商都有关于其产品的可靠性报告，该报告分析了加速寿命测试的结果并提供一个器件故障率。

15.3 混合微电路及故障

混合器件是指有源器件与附着在衬底上并通过导电薄膜相互连接的各种分立元件的组合。混合器件的故障模式类似于其他分立和集成电路，主要是由于制造工艺的缺陷导致的。这种故障无论器件是否工作都将存在。在器件使用之前，各种催化性的环境因素（如高温或振动）都将会激发这种故障的出现。表 15-2 列出了各种催化性的环境因素以及对应出现的故障。

表 15-2 混合微电路——故障催化性环境

催化性环境	环境故障类型	故障模式
机械压力(热冲击、振动)	衬底粘结、衬底破裂、虚焊	开路
高温	损坏电阻	开路或超过容量
热循环	电阻裂开、各种薄膜缺陷、过长焊接时间、薄膜与衬底之间的热扩展系数不匹配	开路或超过容量
高电压、高温测试	各种薄膜缺陷、过长焊接时间、薄膜与衬底之间的热扩展系数不匹配	超过容量
热压和机械压力	模块裂开、引线短接、虚焊、脱焊	开路、短路

15.4 存储IC故障模式

存储IC的故障主要是指无法读取和写入数据、数据存储错误、不能接受的输出电平以及较慢的访问时间等。这些故障产生的具体原因如下所述：

1）开路和短路：单数据位错误可以导致整个器件出现灾难性的故障；

2）译码器开路：导致存储器寻址问题；

3）多路写入：对某一个单元写入时，实际上是对另外一个单元或其他单元的写入（多路写入和寻址惟一性问题）；

4）模式灵敏度：单元内容由于电气上相邻单元的读写操作变成补码（单元干扰、相邻单元干扰、列干扰、相邻列干扰、行干扰、相邻行干扰）。

5）写入恢复：当每个读取周期处于一个或多个写入周期之前，器件的访问时间可能比指定的慢。

6）地址译码器切换时间：这个时间随切换前地址译码器的状态以及正在切换的译码器的状态而变化；

7）读出放大器敏感度：在读取一长串的类似数据位后，存储信息可能是错误的；该数据之后是一个过渡的相反数据数值；

8）休眠故障：存储器在指定的保持时间之前就丢失了存储的数据（例如DRAM）；

9）刷新问题（可能是静态或动态）：刷新过程中，静态刷新测试在器件处于无源状态之后对数据内容进行测试。在动态刷新中，器件仍然可以处于有源状态，而且部分单元被刷新。然后，对所有的存储单元进行测试，以检测没有被刷新的单元中的数据是否还是正确的。

存储器必须进行特定的测试，以确保每个存储位置的完整性。这些功能性的测试主要由4个基本测试组成，分别为：模式测试、背景测试、定时测试和电压测试。表15-3给出了这些测试的概述。

表 15-3　存储器电路的测试

测试类型	内　容
March	一种 N 型测试，用来检测功能单元故障（开路、短路和地址惟一性故障）和数据灵敏度；在该测试中，存储器被全部写入 0，然后对每个存储单元进行测试并以相反的顺序再写入 0，直到第一个地址被读取，之后，以相同的顺序再载入 1，并重复上面的步骤
Masest	一种 N 型测试，用来检测地址惟一性和地址译码问题（开路译码线），这是一种交替多地址选择测试方式，在该方式中，首先在存储器的最小地址单元处写入 0，然后在最大地址单元处写入 0；之后在最小地址 +1 单元处写入 0，在最大地址 -1 单元处写入 0，依此类推；然后，所有的存储单元被读取并进行校验，补码形式的 1 将会被读取到存储器中，然后测试重复进行
刷新	用于测试可接受刷新时间的几种模式之一，通常为 N 型测试
易变对角线	一种 $N^{3/2}$ 型测试，用来检测地址译码器问题、读取放大器故障和读取放大器还原问题，该测试在补码数据域产生一个数据对角线，并随 x 轴变化，产生一个条纹式电极效应
环境干扰	一种 N 型测试，用来检测相邻存储单元的干扰，包括行干扰和列干扰
快速测试	这是一种顺序测试，其中一个基本单元中存储了其他单元的补码数据（例如，基本单元中为 1，而其他单元为 0）
步进式测试	步进式测试类似于快速测试，不同的是存储器中的存储单元会被顺序读取，而不是交替读取
快速/步进式测试：GALPAT 行（跃迁行）行	一种 $N^{3/2}$ 型测试，用来检测相邻单元的步进干扰和行/列干扰
GALPAT 列（跃迁列）步进列	一种 $N^{3/2}$ 型测试，用来检测行/列干扰
GALPAT（跃迁式测试）	一种 $4N^2$ 型测试，用来检测地址译码器问题、相邻存储单元的干扰、行/列干扰和读取放大器还原问题，该测试要耗费很长时间；首先第一个存储单元在初始值为 0 的情况下获取 1 的补码，然后读取存储器中的其他每一个存储单元，这个顺序一直被重复直到所有的存储单元都被测试到，然后，重复进行上面的测试步骤，但是将第一个单元的初始值设为 1，并取 0 的补码
GALDIG（跃迁式对角线）模式测试	该测试用来查找各单元之间不合格的地址跃迁以及在单元对角线中的位置，同时该测试还查找较慢的读取放大器还原情况和由于同一列中单元间的噪声耦合导致的存储数据破坏情况；在该测试中，存储单元中的对角线单元为 1；而其他为 0；所有的存储单元被读取并进行校验，然后对角线被水平移动，上面的步骤一直重复直到对角线穿过存储器，然后，该测试对补码数据进行相同步骤地测试
GALWREC	一种跃迁式写入恢复测试，是一种 N^2 型测试

（续）

测试类型	内　容
WALKPAT（步进式）	一种 N^2 型测试，用来检测地址惟一性和读取放大器恢复问题，该测试要耗费很长时间；首先第一个存储单元在初始值为0的情况下获取1的补码，然后顺序读取存储器中的其他所有存储单元，之后，第一个单元写入0，下一个单元取补码，然后再顺序读取存储器中的其他所有的存储单元；上面的步骤重复进行一遍，然后将所有单元的初始值设为1，并对每一个单元写入0，之后重复前面的步骤（0域中只有一个1，而1域中只有一个0）
双步进式列	一种 $N^{3/2}$ 型测试
MOVI（移动反向）	该测试用于检测地址传输故障，该测试类似于March测试，但该测试在计算顺序中将每一个地址线作为最低有效位使用
扫描测试	每一个单元被顺序写入或读取，寻址的顺序可以从最低地址至最高地址，也可以从最高地址至最低地址，x 轴或 y 轴都可以作为快速轴（分别为行快速轴和列快速轴）
地址补码	寻址顺序中的所有地址线（除了一个）在每个寻址周期内会发生变化
初始测试（可能使用先前的测试结果）测试板 反向测试板	一种 N 型测试，用于检测功能性单元故障和数据灵敏度（对噪声），每个存储单元都要进行检测，以确定能否写入0或1，每个存储单元周围都是存储补码数据的单元（例如，存储逻辑1的单元周围都是存储逻辑0的单元），变量包括双测试板和补码双测试板，两者具有相同的执行时间
伪测试板	该测试板类型由专用的OR数据产生，该数据具有最低有效行地址位和最低有效列地址位（数据XOR X0 XOR Y0），该测试可能不会产生一个真实的元件检测板初始值，该元件具有折叠数字线，因此这些元件就需要一个不同的方程式
0/1	一种 N 型测试。每个存储单元的初始值要么全是1，要么全是0
行带	初始数据中每一行都被它的补码包围着，因此如果某一行中心存储的数据为1，那么其周围的数据就都为0
列带	类似于行带，只是数据是存储在各列中的
双倍变化	初始值的大小是双倍的，两行或两列或一个包含两行和两列的双测试板
奇偶性	初始数据在 x 和 y 的地址奇偶性为奇或偶时产生
位图（可以在系统级水平检测存储器）	用来检测故障单元的模式，描述存储器的工作区域，校验器件的规范；该测试用来修正掩模排列误差、扩散异常、污染缺陷、设计灵敏性和半导体缺陷

存储器的测试无法指定到对每一个单元进行100%的测试，例如一个RAM可能包含 2^N 个不同的数据模式，可以被 N 阶（$N!$）个地址序列寻址而且不使用相同的地址。对于某些高集成度的存储器来说，测试时间可能会很长。例如，对一个4MB DRAM的GALPAT测试可能需要106天的时间才能结束。因此，我们可以看到强制指定器件的测试是不切实际的。任何测试方案必须随不同的存储器而各不相同。

系统级测试可能会采用位图法；在位图中，存储单元被显示在系统的阴极射线管（CRT）显示屏上（例如，故障点会被显示为错误的颜色）。各种不同的图案将被读入到存储器当中，然后对 CRT 进行配置，每个存储单元对应一个像素（图像元素）或者进行压缩，每个像素表示 4、16 或 64 个存储单元，这主要取决于可利用像素的数量以及存储器的容量。缩放特征可以按照不同的比例来显示存储器中对故障区域进行封闭测试的结果。

注意到，在大规模存储器加强系统中，通常采用误差纠错码来纠正硬错误（永久故障）和软错误（例如，阿尔法粒子干扰，即暂时性的干扰）。例如，一个视频 RAM 供应商评估了 1M 单元的软错误率为 3.9FIT（周期时间为 500 ns，V_{CC}为 4.5V）和 4M DRAM 单元的软错误率为 41FIT（可信度为 90%，周期时间为 15.625μs，V_{CC}为 5V）。软错误率主要跟刷新周期时间和工作电压有关，电压越低、周期时间越短，错误率越高。1000FIT 等于 1PPM（即每 10^6 小时出现 1 个错误），1PPM 等于 0.1%/1000（即每 1000 小时出现 0.1% 个错误）。在计算软错误率时，必须添加刷新模式错误率和有源模式错误率，这两个错误率可以从制造商提供的加速曲线中得到。

15.5 IC 封装及故障

器件可以根据封装样式进行分类，封装样式可以是密封的（陶瓷外壳或金属外壳），也可以是非密封的（环氧、硅树脂、酚醛塑料或塑封）。大多数微电路如果保存在干燥低温的环境中，那么其采用的材料产生变化的速度就会非常缓慢。

密封、玻璃、陶瓷或金属都可以有效保护器件免受环境的影响。如果是金属封装，IC 器件在保存 50 ~ 100 年之后，其内部湿度仅仅只有外部空气的 50%。如果是环氧封装，可以保存数分钟至数天。因此，应用于恶劣环境中的设备（例如军事应用），必须采用密封的器件。塑封的器件成本较低，可以抗机械冲击及振动，而且在器件内部不会产生微粒（因为没有空间），该类型的封装适用于小尺寸的器件。但是，所有塑封的器件都会包含一定的湿气，而且具有一定的浸透性（盐性环境也可能影响到硅树脂封装的器件，但是不会影响到线形酚醛环氧封装的器件）。塑封的热扩展系数与冲模的不匹配同样会限制器件的工作温度范围（0 ~ 70℃），而密封器件的工作温度范围为 -55 ~ 125℃。

微电子与计算机技术公司（Austin，Texas）和 Lehigh 大学已经对采用效率更高和更轻的保护外壳来代替密封封装进行了研究。他们的项目（“无密封可靠性（Reliability without Hermeticity，RwoH）”）由美国空军 Wright 实验室发起，并签订了研究合同。RwoH 是一个工业工作组，成立于 1988 年，主要研究非密封

电子封装技术，尤其是为半导体器件和 MCM 提供具有环保性能的封装技术。

电子工程设计发展联合会议（Joint Electron Device Engineering Council，JEDEC）的第 A 112 条规范定义了湿度灵敏性的 6 个级别，第 1 级是最高级别或者称为“无湿度灵敏性”。其他为依次降低的级别；例如，第 3 级保证在 30℃ 的温度和 85% 的相对湿度条件下产品可以保存 168h（例如，防止在低温焊接过程中由于夹杂的湿气出现爆米花式的破裂）；第 4 级可以保证产品在相同的条件下保存 84h；而第 6 级只能保证产品在相同的条件下保存 6h。采用塑封的元器件通常在运输过程中都配备有干燥剂（有效期为 12 个月），而且该元器件只有在使用时才能拆封。超过这个有效期限的元器件，尤其是塑封方块扁平集成电路（Plastic Quad FlatPack，PQFP）封装的器件，必须经过烘焙来清除整个封装内的湿气。夹杂的湿气在快速加热的情况下可以蒸发（例如在回流焊接过程中），产生的气压可能会导致封装破裂。如果在高温和潮湿条件下，污染物会进入 IC 并在一段时间之后产生侵蚀，从而导致故障的发生。

有记载在 1995 年，休斯导弹系统中曾出现过由于可溶于水的有机成分（例如有机酸（Organic Acid，OA）产生溶化，从而导致枝晶长大和短路的故障。我们发现采用高铅玻璃粉（玻璃粉通常用于浸蘸和陶瓷扁平元器件）密封的元器件封装和器件引线上的铅会导致玻璃密封表面的枝晶生长。这些铅来自于高铅氧化物，包括软玻璃自身。高温会加速枝晶的生长，如果在 140℉ 热熔剂条件下，在短短 5min 内就会产生短路。为了阻止这种情况发生，可以采用不同的熔剂并尽量减少 PCB 上元器件内部熔剂的使用。枝晶的生长在 kovar 扁平电路中也出现过，该电路保存在金属条状纸板封装中。产生氯枝晶生长（会导致从引脚到芯片产生严重的漏电）的主要原因是纸板和器件中的化学污染。目前，纤维板和纸板封装已经不再用于器件的封装了。在将引脚直接焊接到引线的元件上还会出现针状结晶的生长。

特定 IC 封装中与应用有关的故障是有限的，分别如下：

1）球状栅格阵列（Ball Grid Array，BGA）或过模片状阵列载体（OverMolded Pad-Array Carrier，OMPAC）：在制造工艺中，这种封装的主要故障是短路（98% 的故障为短路，2% 为开路）；

2）倒装晶片：在制造工艺中，MCM 中倒装晶片的主要故障模式是短路；

3）PQFP：该封装易受到实际应用中潮湿引起的破裂的影响，因此需要回流焊接；

4）方块扁平封装（Quad FlatPack，QFP）：最常见的方块扁平封装制造故障模式是短路（82% 的故障为短路，2% 为开路，16% 为排列误差问题）；

5）带状载体封装（TCP）：带状自动粘结（TAB）封装的主要故障模式为开路（98% 为开路，2% 为排列误差问题）。

表 15-1 中用来检测元器件故障的筛选测试采用的是环境条件，这些元器件可能会由于工艺缺陷而过早出现故障。表中还给出了各种筛选测试方法，这些测试方法可以用来检测各种影响微电路可靠性的缺陷。这些屏蔽可以分各种等级而且成本较低，以便有效区别元器件的各种缺陷，这些等级包括：高温保存测试、温度周期变化测试、热冲击测试、氮气罐体测试（主要测试密封封装性能）以及总泄漏测试（也是测试密封封装性能）。筛选测试一般在元器件存放使用之前进行。合适的 ESD 流程和控制必须严格遵循，而且由于 ESD 退化效应，元器件不能定期进行测试或处理，只能在使用之前进行电气测试。

元器件故障（如引脚与引脚直接的泄漏）也可能由受污染的薄膜引起，这些薄膜处于封装的外密封和嵌入玻璃密封处。SRAM 中的这一类故障模式是由抗静电的 ABS/PVC（丙烯腈-丁二烯-苯乙烯/聚氯乙烯）塑料盘中的硫化铅引起的，塑料盘在快速封装和安装过程中用来存放元器件。硫化铅在封装表面和嵌入玻璃密封上的外观类似光亮的薄膜一样，最终会导致超过 20μA 的漏电（5V 电压时）。采用巯基乙酸盐（或者二丁基锡，或辛基巯基乙酸盐、热稳定剂）材料的 PVC 作为一种添加剂，可以引起采用铅基嵌入玻璃密封元器件的硫基污染（和腐蚀）。因此，PVC 制造工艺正在寻找更加稳定的替代材料，那些性能出现退化的产品必须采用金属盘来保存和处理。

15.6　除铅

减少环境中铅的含量关系到人的健康和安全，这一点是目前关注的热点。一个典型的微处理器中含有约 0.2g 的铅，而且计算机的主板中包含 2～3g 的铅，因此整个美国的电子电路市场就使用了全球约 2% 的铅产品；出于环境的考虑，人们提出了各种环保规定，这些规定限制了铅和其他有害物质的使用，这些规定由欧盟在欧洲议会和欧盟委员会的草案上提出，如下所述。

1）从 2000 年 10 月起大规模地执行 ELV（End of Life Vehicles）草案 2000/53/EC：该草案要求产品中不能包含重金属，如汞、镉、六价铬以及铅。该草案还要求汽车制造商在 2002 年 7 月 1 日以后生产的汽车中必须建立再循环系统，而在 2007 年 7 月 1 日之后所有的汽车必须建立再循环系统。铅仍然可以在铜制品和可软焊的产品中用作合金添加剂。

2）WEEE 草案即损耗电气和电子设备（Waste Electrical and Electronic Equipment，WEEE）草案 2002/96/EC：该草案扩充了 ELV 草案关于再循环的要求，包含了广泛使用的电气和电子产品和设备。WEEE 从 2003 年 2 月 13 日起生效，并在 2004 年 8 月 13 日成为了欧洲的国家法律条款，从 2005 年 8 月 13 日起应用到消费类产品当中。但是其中条款 2（3）也这样描述：“那些关系到保护成员国

家、武器、军需品和战争材料基本利益的装备不受本草案限制。但是这一点不适用那些不准备用于特殊军事用途的产品。”

3）RoHS 草案：即关于对电子和电气设备中有害物质的限制，该 2002/95/EC 草案建立了电子和电气设备中有害物质含量的标准和限制条款。该草案从 2003 年 2 月 13 日起生效，并在 2004 年 8 月 13 日成为欧洲的国家法律条款，从 2006 年 7 月 1 日起广泛应用到各类产品当中。草案中禁止或严格限制的物资包括：铅、汞、镉、六价铬、某些溴化阻燃剂（PBB）以及多溴化二苯醚（PBDE）。

在电路板装配中推荐使用的无铅焊料配方为 Sn-Ag-Cu，不过也还有其他类型的焊料，如镍-钯（NiPd）和镀金的镍-钯（NiPdAu）。采用低温除铅处理（215℃时 50%/50% 的锡/铅配方和 230℃时 40%/60% 的锡/铅配方）的有源元器件和温度高达 260℃的无铅处理过程采用不光滑的纯锡作为焊料。焊料中使用铅的原因是基于多种潜在可靠性因素的考虑。纯锡的器件引线已经证明会导致电子器件边界处出现相互的金属迁移和针状锡结晶体，这种锡结晶体会导致短路故障发生（这就是为什么军用产品不能遵循草案的原因）。

国家电子工艺机构（NEMI）已经关注到了无铅设备中“针状锡结晶体”的问题。该机构将针状锡结晶体定义为：单晶锡上，从电镀表面自动散发形成的柱状或圆柱形细丝（极少出现分枝）。针状锡结晶体具有以下特性：

1）纵横比（长度/宽度）大于 2；

2）可以纽结、弯曲以及扭曲；

3）通常具有一致的横截面形状；

4）可能有条痕或环。

推荐的测试方法是：

1）温度周期（-55～85℃，约 3 个周期/h）；

2）温湿度测试条件为 60℃/93% RH；

3）保存环境（空调器）。

注意：如果镀锡的元件在低于 13℃的温度下保存超过 1 个星期，锡就会发生相变，变成粉状形式（称为“锡瘟”）。

15.7 筛选测试及再筛选测试

在 20 世纪 70 年代，美国海军颁布了一项政策，要求接收到的所有元器件必须进行再筛选测试（由承包商或独立测试实验室完成），这是因为有证据表明联合陆军海军（Joint Army Navy，JAN）鉴定的元器件（包括 JANTX、JANTXV 分立元器件）没有达到军用硬件的使用质量标准（DoD 草案 4245.7-M 关于产品开

发至产品成形的过渡过程，该草案签署于1984年，至1986年得到广泛应用）。

再筛选测试通常包含以下几部分：

1）破坏性物理分析（Destructive Physical Analysis，DPA）检查测试：该测试中，在每一批元件中挑选两片（最少），将这两片元件切开，并检查元件的工艺水平。如果发现元件的工艺水平很差，那么这一批元件就可以拒绝使用（这一批元件是从某一制造商的装配线上随机抽选的）。

2）微粒影响噪声探测（Particle Impact Noise Detection，PIND）测试：该测试中，对每个混合元件或具有内腔的IC进行振动，其中IC的冲模是玻璃隔绝的（通过玻璃外层隔离，图15-5给出了采用此类玻璃外层上有裂缝的IC例子），然后附着在元件上的传感器就可以非常准确地检测出元件内部周围是否存在微粒。具有内部活动块的元件是不允许出货的。

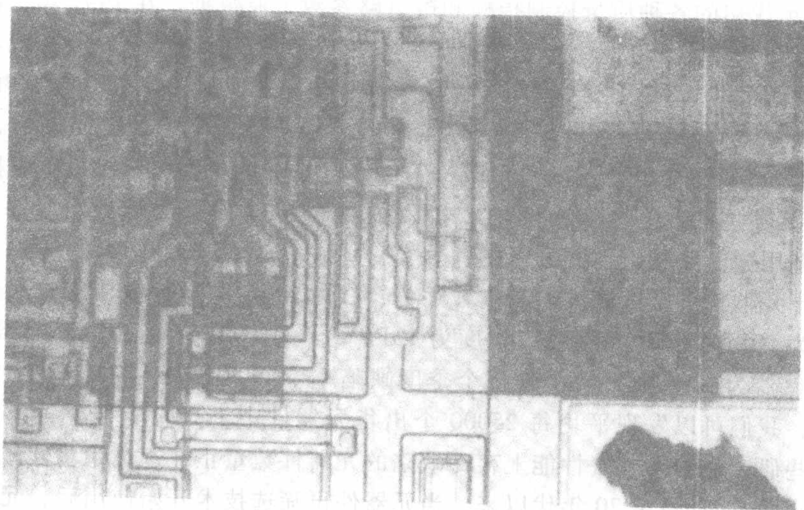


图 15-5 冲模玻璃外层具有裂缝的IC

不间断（Go-on go）电气测试、静态测试和动态测试要求分别在低温、室温（25℃）和高温条件下进行。密封性测试要求测试密封的完整性。

为了找到进行元器件再筛选测试的原因，我们必须调查那个年代的测试数据。在1981年，有报道称位于美国印第安那州Crane的海军武器支持中心（NWSC）发现IC中元器件的缺陷率高达15%，分立半导体的缺陷率高达17%。在1983年1月，有报道称一个独立测试实验室发现线性IC的不合格率为16.7%，数字IC的不合格率为8.4%，CMOS IC的不合格率为9.2%；到1984年10月，该实验室发现线性IC的缺陷率为5.5%，数字IC的缺陷率为3.7%，晶体管的缺陷率为8.2%，光电子元器件的缺陷率为3.8%。在1983年，美国国

防部电子中心发现非军用规范的元器件中半导体缺陷率为8%，略高于1%的军用规范元器件缺陷率。1986年，一个军用产品提供商评估发现IC的再屏蔽故障率只有0.9%，而晶体管的再屏蔽故障率只有1.5%。1989年，美国海军的一个专用计算机工作站制造商公布了以下再筛选结果：IC/半导体测试：8312个（127种元器件）元器件；其中，总共不合格的产品有25件——PIND/DPA：（16个混合元器件PIND故障、6个电气故障、3个机械故障——引脚破裂），不合格率为0.30%。

详细分析混合元器件的故障，在99个经过测试的军用合格振荡器中，6个被驳回，不合格率为6.1%，53个非军用合格振荡器经过测试后，有10个被驳回，不合格率为18.8%。

美国的半导体工业指导了一个质量统计计划，用来监控和报告1985年至20世纪90年代初的各种质量控制指标和微电路参数工业数据。在1991年，数据分析显示，在平均每10000个出货的元器件中，只有一个存在电气缺陷，或者称缺陷率为100PPM（百万分之一百）。这些数据是按照JEDEC的标准16公布的，即每百万个元器件的微电路出厂质量等级评估指标，而且这些数据代表了JAN合格元器件、DESC制图元器件、MIL-STD-883C标准屏蔽元器件以及源控制图（SCD）元器件的数据。上面的抽查过程抽查了超过150百万个元器件，其中80%的军用微电路是由地方公司提供。

基于元器件类型的缺陷率具体数字为：线性元器件为100PPM；双极性数字元器件为50PPM（1991年第一季度为10PPM）；MOS数字元器件为<100PPM；MOS存储器为<160PPM（平均每个季度缺陷率下降30PPM）。如果将所有的数据合并，我们可以发现平均每25000个出货元器件中，只有1个元器件存在缺陷。这些研究表明在电气性能上存在缺陷的元器件数量正在逐渐稳步地减少。这个分析表明自20世纪70年代以来，当元器件再筛选技术开始使用后，元器件的质量得到了重要发展，而且再筛选技术可能只在超高可靠性的产品中才需要使用（如太空产品）。

半导体、集成电路和混合元器件筛选测试

表15-1归纳了各种评比测试方法，这些测试方法可以用来检测各种半导体故障机制。其中，与应用环境无关的故障（氧化缺陷或扩散缺陷）在元器件的工作过程中会加剧（例如预烧测试）。

与应用环境相关的故障（焊接缺陷或电镀缺陷）会随温度和机械压力（如振动）的变化而加剧。

与时间和环境相关的故障（金属迁移、侵蚀、由于不同类金属混用而导致的金属互化物）在元器件的工作过程（例如更高温度的预烧测试）中会加剧。

15.8 静电放电效应

IC 非常容易受到静电放电的影响，静电不用达到产生火花的地步就已经足够对 IC 产生破坏了。图 15-6 给出了一个 D/A 转换器电路中 ESD 故障的例子。图 15-7 给出了金属镀层下进行扩展时的钝化故障例子。

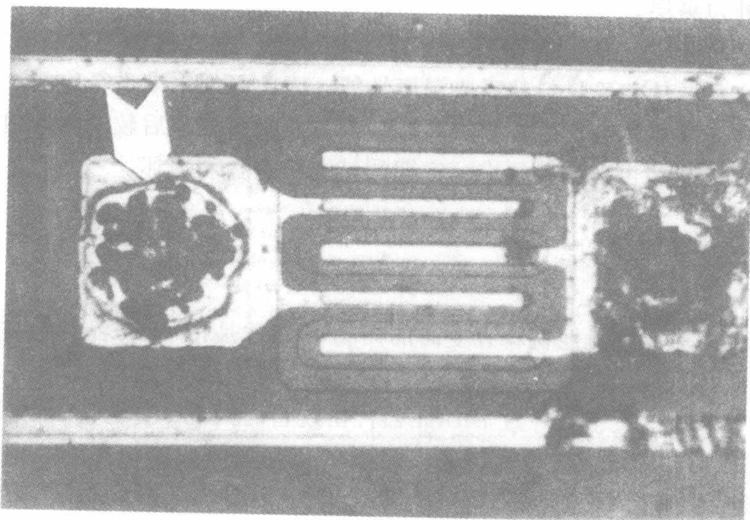


图 15-6 D/A 转换器电路中的 ESD 故障



图 15-7 金属镀层下进行扩展时的钝化故障

为了防止产生ESD破坏，我们可以采用各种不同的抗静电材料。典型的抗静电材料（用来包裹某些塑封IC汲取运输管）包括清洁物质，这些清洁物质用来清洁湿润IC的表面。外包裹层的性能与其湿度有关，在高湿度条件下性能更好。但是这些外包裹层会被耗尽，因此需要利用静电探测器定期对加工后的元器件表面进行静电检测。

加工后的IC汲取管在使用后必须重新进行加工，因为IC引脚在滑过汲取管时会刮掉外包裹层。

粉色的聚合物包裹层也可用于元器件的封装。这些包裹层也可能会失去效能，因为它主要依赖于材料的吸收湿度性能，因此必须定期进行检测。

镀镍包裹层可能会产生传导中断，因为重复封装处理后较薄的外包裹层会出现破裂，因此这些包裹层必须进行定期测试并检查物理损坏。

名词解释

阿尔法粒子：放射性材料衰变时的产物（通常是指IC陶瓷封装材料中辐射出的放射性射线）。这种粒子带正电，其电量等同于两个电子的电量，而且辐射速度极快。阿尔法粒子会导致DRAM器件出现暂时混乱，称为“软错误率”。

焊接区：IC冲模上镀金的区域，该区域允许将导线或电路元器件连接到冲模上（类似于导线连接）。

预烧（Burn-in）：元器件测试方法。在较高的电压和温度条件下，经过一定的时间，该测试可以筛选出初始无效故障（缺陷或较差的元器件）。

陶瓷：一种无机物，非金属粘土或玻璃状的材料。其最后的性能是在高温下形成的。

DC测试：一种测试方法，用来测量静态参数，如泄漏电流。

美国国防部电子支持中心（Defense Electric Supply Center，DESC）元器件：用来表示由DESC开发的标准军用制图元器件，DESC位于美国俄亥俄州的哥伦布市（以前位于Dayton）。

冲模：一种集成电路，与芯片齐名。

冲模信息交换（Die Information Exchange，DIE）：由芯片制造商和软件提供商提出的一种规范，提供了标准形式的基本冲模信息。

DIE粘接：将IC芯片（或DIE）附着在衬底或顶部。

DIE分离：实际IC芯片从封装内分离出来的过程。

破坏性物理分析（DPA）：将元器件切开并分析工艺的完整性和工艺水平。

静电放电（ESD）：积累在非导体上的电荷瞬间转移到导体上或接地的过程。

故障率（Failure In Time，FIT）：等于10^9h内发生故障的次数。

玻璃料：一种软化熔点相对较低的玻璃制品。

功能测试：通过真值表来检测元器件正常工作的方法。

密封：对元器件进行封装达到不透气的程度（通常氦的含量低于 1×10^{-6}cc/s）。

联合陆军海军（JAN）标准：当 IC 中的元器件完全符合 IC MIL-M-38510 要求（现在由 MIL-M-38535 代替）和半导体 MIL-S-19500 要求时使用。标准军用制图（SMD）计划的 JAN B 级集成电路标准是武器系统设计的推荐标准。

JANTX：一种前缀，用来说明军用规范元器件接受了额外的筛选测试，例如 100% 的 168 小时预烧测试。

JANTXV：一种 JANTX 元器件，附加了预封装可视性要求。

联合电子元器件工程委员会（JEDEC）：电子工业协会（EIA）的一个部门。

引线：一种导电路径，通常为电子元器件中用来连接外部电路的独立部分。

引线框：元器件封装中的金属部分，用来实现冲模与其他电路之间的电气连接。

掩模：电路单元中的模板，通过该模板将电路图案影射到芯片（冲模）的感光涂层上。影射的区域会被揭掉，剩下一个电路图案。

电镀：将很薄的金属包裹层沉淀在 IC 或半导体上。

钝化：在冲模表面形成一个隔离的绝缘层。钝化通常是通过硅的热氧化实现，并形成一层很薄的二氧化硅（PECVD 氧化物和 PECVD 氮化物在稍低的温度下（低于 450℃）的合成物）。也可使用其他钝化绝缘层，例如硅玻璃（硅氧氮化物）。

微粒影响噪声探测测试（PIND）：一种测试方法，在该方法中，对空腔元器件进行振动和监听，通过材料的噪声来确定元器件封装内部是否存在疏散的微粒。这些疏散的微粒可能具有导电性（例如来自金共晶冲模焊接操作过程的金制薄片微粒），会造成短路。这种测试在军用 B 级器件中不是 100% 都需要进行，而 S 级的所有元器件或航天器件都需要进行此类测试。

爆米花式破裂：一种塑封破裂，或者由相位改变和扩展导致的脱层。

百万分之几（Parts Per Million，PPM）：每百万个元器件中故障元器件的数量。缺陷元器件的统计评估指标，通常可信度高达 90%。

Schmoo 图：一种 X-Y 图，该图给出了当 X 和 Y 坐标参数发生变化时，指定测试的合格区域和不合格区域。

软错误：元器件输出值中出现的一种错误或混乱现象（通常表现为存储器件中的单数据位输出错误），该错误只是暂时的。

衬底：一种支撑材料，在其上面可以制造集成电路（IC）；或者是混合材料，IC（或其他元器件）附着其上。

针状锡结晶体：在电镀层表面形成的一种针状单晶体生长情况。

导线连接：半导体冲模焊接区域和引线框或接线端之间的导线连接。

感谢

在此，要感谢 ISSI 公司（Integrated Silicon Solution Inc.）的 Ron Kalakuntla 和技术销售小组的 Lawrence Taccour 对本章的审阅。

参考文献

[1] Arnold, H. D. 1981. Institute of Environmental Sciences Proceedings. ESSEH (Environmental Stress Screening of Electronic Hardware), Institute of Environmental Sciences (Mt. Prospect, IL) Conf., Sept. 21.

[2] Bley, W. 1981. Testing high-speed bipolar memories. Fairchild Camera and Instrument Corp., Mountain View, CA, Nov.

[3] CALCE (Computer Aided Life Cycle Engineering) News, Jan. 1993. Workshop On Temperature Effects, University of Maryland, College Park, MD.

[4] Chester, M. 1986. DOD's rescreening of chips stirs controversy. *Electronic Products*, Oct. 15, pp. 83-83.

[5] Costlow, T. 1995. MCM substrates mixed. *Electronic Engineering Times*, Jan. 16.

[6] Ellis, M. 1984. Non-Mil Defects Surpass Mil. *Electronic Buyers' News*, Nov. 26, p. 6.

[7] Frye, M. A. 1994. Extension of the implementation date of MIL-I-38535 and MIL-S-19500 regarding the prohibition if pure tin as a plating material. Defense Logistics Agency, Defense Electronics Supply Center, Dayton, OH (Letter dated Jan. 21).

[8] GIDEP. Government and Industry Data Exchange Program (Corona, CA), Problem Advisory, 7G-P-95-01 (Jan. 1995), Microcircuits, Flux, Soldering, Liquid; Problem Advisory G4-P-93-01, Packaging, Cardboard, Flatpack, Microcircuit; Problem Advisory ZW-P-9301A (Jan. 1993), Storage Container, Material, Contamination; Problem Advisory S4-P-93-01 (Jan. 1993), Transistor, Tin Plating, Whisker Growth.

[9] Gulley, D. W. 1992. Texas Instruments, mean time between events, a discussion of device failures in DRAMs. 9 (5), Sept.

[10] Hamilton, H. E. 1984. Electronics test. Micro Control Co. Minneapolis, MN, April.

[11] Hitachi. 1991. Reliability report, Multiport Video RAM. HM534251, MC-883, Hitachi, Ltd., Tokyo, Japan, Oct. 14.

[12] Hnatek, E. R. 1983. The case for component rescreening. *Test &Measurement World*, Jan., pp. 18-24.

[13] Hnatek, E. R. 1984. ICs for military and aerospace show dramatic jump in quality, reliability, military/space electronics design. Viking Labs., Inc. *Military/Space Electronics Design* (McGraw-Hill), Oct., pp. 27-30.

[14] Hu, J. M., Barker, D., Dasgupta, A., and Arora, A. 1993. Role of failure-mechanism identification in accelerated testing. *Journal of the IES* (July/Aug.), Institute of Science (Mt. Prospect, IL).

[15] ESSEH (Environmental Stress Screening of Electronic Hardware) 1985. Environmental stress screening for parts. Institute of Environmental Sciences Proceedings, *ESSEH*, Sept.

[16] Klinger, D. J., Nakada, Y., and Memendez, M. A. *AT&T Reliability Manual*, Van Nostrand Reinhold, 1990.

[17] Lawrence, J. D. Jr. 1983. *Parallel testing of memory devices*. Reliability Inc., Houston, TX, Oct.

[18] RADC. 1988. *Reliability/maintainability/testability design for dormancy*. Lockheed Electronics Co., Rome Air Development Center, Rept. RADC-TR-88-110, May. (Available from the Defense Technical Information Center (document#AD-A202-704)), Defense Logistics Agency, Department of Defense.

[19] Meeldijk, V. 1990. *Electronic Components: Selection and Application Guidelines*. Wiley Interscience, New York, Chap. 10 and 11.

[20] Meeldijk, V. 1990. Effects of storage and dormancy on components. *Electronic Servicing and Technology Magazine*, Dec., pp. 6-11.

[21] Micron Technology, Inc., Boise, ID. 1991. *Quality/Reliability Handbook*, 4/91, Reliability Monitor, 4 M DRAM, rev. 10/91 and 1 MEG DRAM book.

[22] MIL-HDBK. 1988. *Electronic Reliability Design*. MIL-HDBK-883 Military Handbook, Oct. 12.

[23] MIL-STD. 1995. NASA standard electrical, electronic, and electromechanical (EEE) parts list. MIL-STD-975, March 17.

[24] MIL-STD. 1992. Electronic parts, materials, and processes for space and launch vehicles. MIL-STD-1547, Dec. 1.

[25] MIL-STD-19500. 1994. General specification for semiconductor devices, April 15.

[26] Motorola, Phoenix, AZ. 1982. Dynamic RAM quality and reliability report. MCM6664A/6665A 64K.

[27] Murray, J. 1994. MCMs pose many production problems. *Electronic Engineering Times*, June 11.

[28] Navsea Systems Command. 1991. Washington, D. C. *Parts application and reliability information manual for Navy electronic equipment*. TE000-AB-GTP-010, Navsea Systems Command (stock number 0910-LP-494-5300). March.

[29] O' Connor, P. T. D. 1985. *Practical Reliability Engineering*, 2nd ed. Wiley, New York.

[30] RAC. 1991. *Failure mode/mechanism distributions*. FMD-91, Reliability Analysis Center, Rome, NY.

[31] School, R. 1985. *Effective screening techniques for dynamic RAMs*. Pacific Reliability Corp. presented at Electro., Institute of Electrical and Electronic Engineers.

[32] Scharfer, S. 1994. *DRAM soft error rate calculations*. Design Line, 3 (1).

[33] Semiconductor Industry Association (San Jose, CA). 1991. *Quality statistics report on military integrated circuits*. Government Procurement Committee.

[34] Smith, D. J. 1985. *Reliability and Maintainability in Perspective*, 2nd ed. Halsted Press, Wiley, New York.

[35] Somos, I. L., Eriksson, L. O., and Tobin, W. H. 1986. Understanding di/dt ratings and life expectancy for thyristors, *PCIM Magazine*, Feb.

[36] U. S. Army Material Command AMCP-706-196.

[37] Willoughby, W. J. Jr. 1980. Military electronics/countermeasures: View from the top (interview), Aug., pp. 14-20, 60-61.

[38] Yalamanchili, P., Gannamani. R., Munamarty, R., McClusky, P., and Christou, A. 1995. Optimum processing prevents PQFP popcorning. CALCE Electronic Packaging Research Center, MD, *SMT*, May.

备注

关于故障模式的更多信息，读者可以参考如下资料：

[1] The Reliability Analysis Center (RAC), Rome, NY, Failure Mode/Mechanism Distributions, 1991, FMD-91.

[2] AT&T Reliability Manual, by David J. Klinger, Yoshinao Nakada, and Maria A. Memendez, published by Van Nostrand Reinhold, 1990.

第16章　基本计算机体系结构

Joy S. Shetler

16.1　引言

计算机体系结构的设计过程是变化多样的，也是非常复杂的。每一种新的体系结构都必须实现一系列的目标，并努力在计算机应用中开创出一片新的领域。任何一个计算机系统都是由一个或多个处理器构成的。在多处理器和单处理器系统中，将会涉及到各种计算机概念。很多从事计算机研究的人总结出一个结论即计算速度和吞吐量将会如同过去一样发展，两者将齐头并进，而不会过重依赖于技术的创新。但是，具体的实现方式仍然是任何计算机系统必须考虑的重点之一。

16.2　计算机体系结构的定义

下面给出了计算机系统的一些重要属性：

1）控制路径结构；

2）数据路径结构；

3）存储器组织机制；

4）应用技术；

5）时钟模块的数量（单模块和双模块）；

6）时钟速度；

7）各单元流水线操作等级；

8）系统的基本拓扑结构；

9）控制路径、数据路径和互联网络的并行等级。

在某些情况下，一条指令的执行算法在数据流结构或脉动（Systolic）阵列中同样重要。上述这些属性与其他属性密切相关，而且会随系统的实现方式变化而变化。例如，相关的时钟速度可能会取决于采用的技术和流水线操作等级。随着更为复杂多变的系统的开发，计算机体系结构的分类也变得越来越复杂了。最常见到分类方案为“Flynn 分类法”，该分类法定义了4种不同的基本计算机体系结构类型，分别为：单指令单数据（Single Instruction Single Data，SISD）结

构、单指令多数据（Single Instruction Multiple Data，SIMD）结构、多指令单数据（Multiple Instruction Single Data，MISD）结构以及多指令多数据（Multiple Instruction Multiple Data，MIMD）结构。具体的分类是根据在控制或数据路径中存在单个或多个路径来区分的。该分类法的好处是应用起来非常简单实用，该方法定义的计算机体系结构优势在实际应用中也得到了很好证明。业界同时也提出了其他的分类方法，这些方法更加详细地定义了最新计算机结构中的并行机制。

16.3 单处理器系统

如图 16-1 所示，一个单处理器系统通常由 1 个控制器、1 条数据路径、1 个存储器和 1 个输入输出（I/O）单元构成。这些单元的功能可以合并，而且每个单元都有很多名称。合并的控制和数据路径有时称为“中央处理单元（CPU）”，数据路径也称为“算术逻辑单元（Arithmetic Logical Unit，ALU）”。

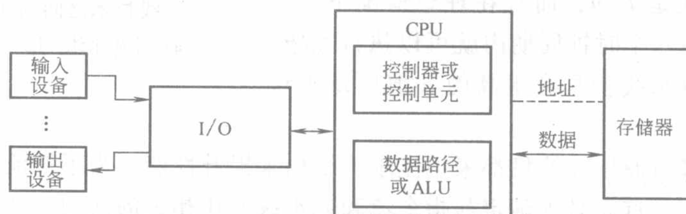


图 16-1　单处理器系统的功能模块框图

处理器中可以通过流水线和多功能单元来实现硬件并行机制，如图 16-2 所示。这种并行机制允许指令在执行过程中按照时间进行交叠，因此，在流水线系

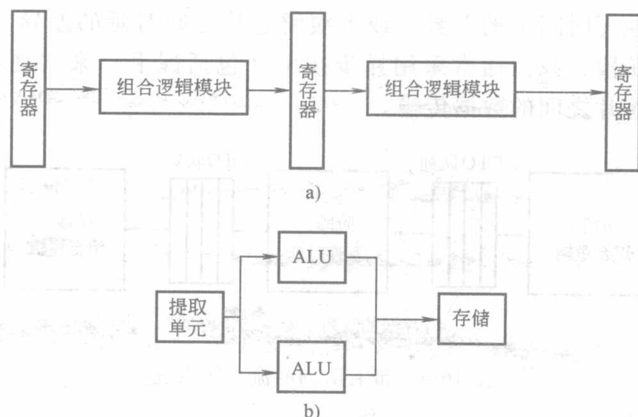
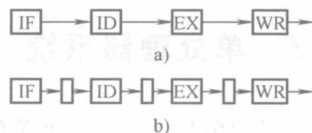


图 16-2　处理器系统中的组合并行方式
a）流水线　b）多个单元

统或多功能单元系统的独立功能单元中，相同指令流或线程中的多个指令可以在每一个流水线上交叠执行。“上下文”定义了一个处理器在执行处理过程中所需的信息，包括程序计数器或指令计数器、堆栈指针、指令寄存器、数据寄存器或通用寄存器的内容。指令、数据和通用寄存器有时也称为“寄存器组”。上下文必须保存在处理器中，以使执行时间最小化。有些系统允许单指令流中的多条指令同时被激活，这就需要大量的上级硬件来支持，而且必须对每条指令流的各条独立指令进行监控。



缩略语说明
IF－取指令　ID－指令译码
EX－执行　WR－写结果

图 16-3　基本处理器流水线技术之间的比较
a）非流水线　b）流水线

流水线系统是在 20 世纪 70 年代流行起来的。在流水线系统中，指令的执行操作过程被分解成多个阶段，如图 16-3a 所示。在流水线中，有多条指令被同时执行，而每条指令处于不同的执行阶段。如果通过流水线的总时延为 D，而且流水线被分解为 n 个阶段，那么最小的时钟周期就是 D/n；而且在理想情况下，一条新的指令在一个时钟周期内就可以执行完毕。更为复杂的流水线就具有更高的 n 值以及更快的时钟周期。

目前大多数商用计算机都采用流水线结构来提升性能。为了使流水线中的时钟周期最小化，首先必须确定与指令流和流水线操作相关的问题，并通过技术改进（如指令和数据预取、编译器技术、数据或指令高速缓冲）来解决发现的问题，业界对此进行了大量的研究。通过流水线每个阶段的时延还与流水线每个阶段的逻辑复杂程度有关。在很多例子中，流水线每个阶段逻辑单元的实际流水线时延远比理想值（D/n）大。我们可以通过在流水线各个阶段添加队列来缓解合并逻辑单元在执行时间上的差异，或者缓解芯片之间时延的差异（见图 16-4）。在流水线各个阶段，我们通常采用异步技术（包括握手）来实现不同时钟条件下逻辑单元或芯片之间的数据传输。



图 16-4　带有队列的流水线系统

我们知道，计算机的硬件设计越简单越好。那些将逻辑功能的数量最小化的系统更加容易设计、测试和调试，而且功耗也更小，运行速度也更快（工作在

更高的时钟速率)。满足上述条件的计算机体系结构有两种,分别为"精简指令集计算机(RISC)结构"和"SIMD计算机结构"。RISC体系结构通常用于精简增加的代码长度,并预取下一条指令,以获得更快的时钟速度和更低的指令集复杂度;SIMD体系结构(在多处理器系统部分描述)利用单指令流来在成千上万个并行工作的处理器上处理超大规模的数据集。

在20世纪80年代初,RISC处理器为小型处理器的发展取得了突破,并且一直支配着小型计算机的发展,如本书所述。计算机的性能 p 可以通过下面的关系式来描述:

$$p = \left(\frac{\text{计算量}}{\text{指令}}\right)\left(\frac{\text{指令数}}{\text{时钟周期}}\right)\left(\frac{\text{时钟周期数}}{\text{秒}}\right)$$

上面关系式中的第一部分(计算量/指令)是指被执行指令的复杂度,随着处理器的结构和当前执行指令的类型变化而变化;第二部分(指令数/时钟周期)的倒数通常在单处理器设计中作为"每条指令所需时钟周期数(CPI)";最后一部分(时钟周期数/秒)通常可作为处理器的时钟周期。在RISC处理器中,只提供了最常用的硬件,这样就精简了每条指令所占用的计算量,并摒弃了复杂的指令。在编译时,有很多简单基本指令可以用来执行大部分的操作,而这些操作之前是由复杂指令执行的。因此,RISC处理器可以比复杂指令集结构(CISC)的处理器执行更多的指令。通过简化硬件,时钟周期也会减小。如果执行更多指令时的时延(RISC设计)比采用增加时钟周期的方法来执行所有指令(CISC设计)的时延小,那么整个系统性能就被提升了。而更先进的编译器设计技术和大规模片上高速缓冲技术也进一步推动了RISC设计性能上的发展(Hennessy and Jouppi, 1991)。

RISC体系结构的效率比传统的CISC体系结构效率更高的原因之一是采用了大规模片上高速缓冲和寄存器组。由于参考效应(存储器分级体系中有介绍)的位置决定了大多数指令和数据参考值的性能,而片上高速缓冲器和大量的寄存器组可以减少指令的数量以及每条指令执行之前提取数据的数量。大多数RISC处理器采用流水线来交叠指令的执行过程,从而减小时钟周期。编译器技术可以用来在顺序执行的程序中实现自然的内在并行性。

寄存器窗口是寄存器的一个子集,用于处理特殊指令。这些寄存器窗口可以是输入寄存器、输出寄存器,也可以是用于处理指令的临时寄存器。一个指令集的输出值可以作为另一个指令集寄存器窗口的输入值。这种技术提高了寄存器的使用效率,并提高了某些体系结构中流水线的等级。

将这些RISC概念应用到大规模并行处理系统中时,将会遇到很多挑战。规模越大,基于RISC并行处理的系统产生的问题越复杂,而且通信和资源位置会严重影响到系统资源的利用率。除非系统中合并了特殊的路由芯片,否则处理器

就需要花费大量的时间来处理其他处理器的要求或者数据和指令等待问题。采用大量的单 RISC 处理器就意味着必须对高速缓冲的访问进行控制（监听高速缓冲）或严格限制（目录高速缓冲），以维持整个系统中数据的连贯性。

RISC 体系结构很难处理的问题包括那些产生大量分歧的问题以及利用超大规模数据集时产生的问题。在这些问题中，大量的指令或数据高速缓冲器的使用率很低，而且从存储器中提取大量新指令或数据的提前时间量远高于时钟周期，因此会造成处理器处于“饥饿”状态，而在处理大规模数据集时编译器技术的使用也会受到一定的限制。

如果单指令流的速度更快，那么流水线处理器就必须增加流水线的深度，或者拓宽流水线的数据路径或控制路径。后者可以通过在每个时钟周期内发送多条指令或超标量体系结构（Superscalar）来实现，也可以通过超长指令字（VLIW）结构来实现。在 VLIW 结构中，很多操作都是在一个单指令中并行执行的。一些研究人员提出了超标量体系和超流水线（Superpipeline）之间正交关系的概念（Hennessy and Jouppi，1991）。在超流水线设计中，流水线的深度从基本流水线中得到了拓展；而在超标量体系设计中，流水线的水平宽度得到了拓展（见图 16-5）。

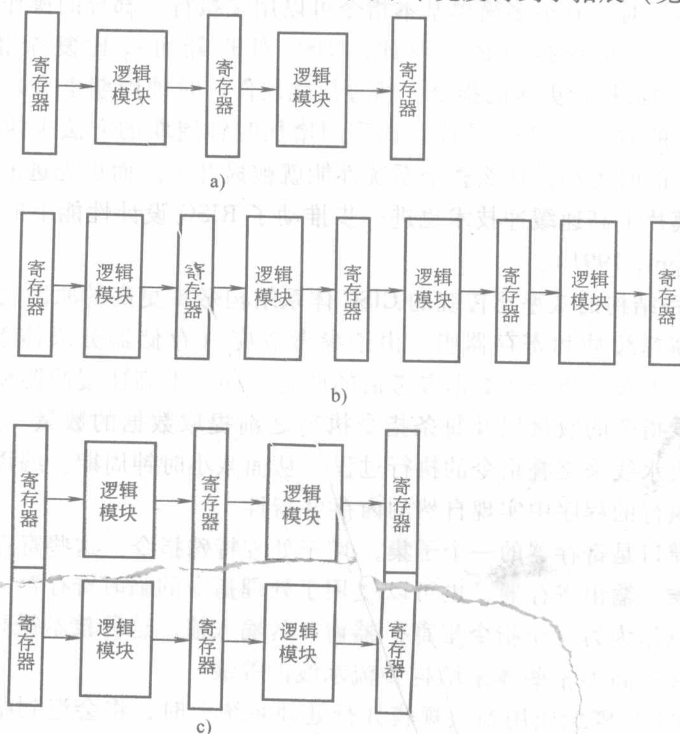


图 16-5　超标量体系和超流水线系统

a）流水线　b）超流水线　c）超标量体系

为了实现整体性能上的提升，由超流水线带来的速度剧增必须伴随有很高的资源利用率。空闲资源对系统的性能来说是毫无意义的，反而会增加系统的总体成本和功耗。随着流水线深度的增加，单指令流无法使处理器中的各个流水线阶段得到充分利用。指令流中的控制和数据依赖性限制了给定指令流中可被激活指令的数量。空操作（NoOp）或空指令被插入到流水线中时，就会产生“气泡”。由于 NoOp 不执行任何操作，因此处理器的周期就被浪费了。为了提高流水线的利用率，我们可以采用各种技术来使存储器的访问提前量降到最小（如预提取一定数量的指令或数据、软件流水线、曲线规划、折叠分析以及寄存器重命名）。这种高效率的并行机制会产生一个我们不愿看到的结果，即某些预提取的指令或数据可能无法使用，这样就会导致在提取无效信息时寄存器的宽度利用率变得很低。在单处理器系统中，这种情况是可以接受的；但是在多处理器系统中，随着寄存器访问次数的增加，这种情况会降低系统的整体性能。

超标量体系处理器采用多指令发送逻辑模块来使处理器一直处于繁忙的工作状态。实际上，在每个时钟周期内，单指令流中只发送 2 条或 3 条指令，这样可以拓展处理器中的控制路径和部分数据路径。VLIW 处理器可以采用不同的功能模块来并行执行多个操作。每条指令由多个操作域构成，结构非常复杂。这些技术的利用率取决于编译器对指令流中指令的分割，或者取决于运行过程中通过构建额外的硬件来完成的分割操作。另外，这些技术还会受到单个指令流内在并行机制数量的限制。

完全使用流水线的方法之一就是使用独立指令流中的指令或线程（线程是指并行结构指定的一段代码的执行过程）。部分处理器允许在程序中包含多个线程。一个线程可以看做是一个工作单元，该工作单元可以由程序设计师定义，也可以由并行编译器定义。在执行过程中，一个线程可以产生代码并行执行过程所需的其他线程。多线程可以降低单处理器系统中的存储器长响应时间效应。当存储器系统的服务缓冲器错过一个线程之后，处理器可以执行另一个线程。多线程也可以拓展到多处理器系统中，并允许 CPU、网络和存储器同时使用。

为了从多线程硬件中获取最好的性能，需要一个兼容的软件环境，而新的计算机语言和操作系统提供了这些环境（Anderson，Lazowska and Levy，1989）。多线程结构充分利用了这些优势来提高系统性能。

16.4 多处理器系统

计算机系统中采用的并行机制包含了不同的等级，最简单的等级就是在系统中只包含几个处理器。如果采用的并行机制是处理器等级，那么通常采用的体系结构就是下面三种之一：SIMD、MIMD 或多计算机系统结构。

SIMD结构允许在同一条指令流中同时执行多条数据流，如图16-6a所示。但是SIMD结构也存在一些问题，因此SIMD结构采用单指令流来避免合并多指令流时存在的缺陷。通常，这种结构要求数据为位序列，因此该结构得到了广泛应用，如计算机制图和图像处理；在这些应用中，SIMD结构可以提供重要的数据处理能力。对于那些要求单数据流必须由单指令流过程的应用来说，SIMD结构的工作速度比其他类型的结构慢，因为每次只有一条指令流被激活。为了克服这种结构上的缺陷，人们提出了"组合SIMD/MIMD"结构。在SIMD系统中，一条指令流可以控制成千上万条数据流，每一个对所有数据流的操作都可以同时执行。

MIMD系统允许多个处理过程共享一组常见的处理器和其他资源，如图16-6b所示。在协同操作的环境中，多个处理器被联合在一起共同执行各种程序。例如，每次一个处理器执行一个处理过程。传统MIMD体系结构的难点在于，当指令流停止时，资源会被完全占用（由于数据依赖性、控制依赖性、同步问题、

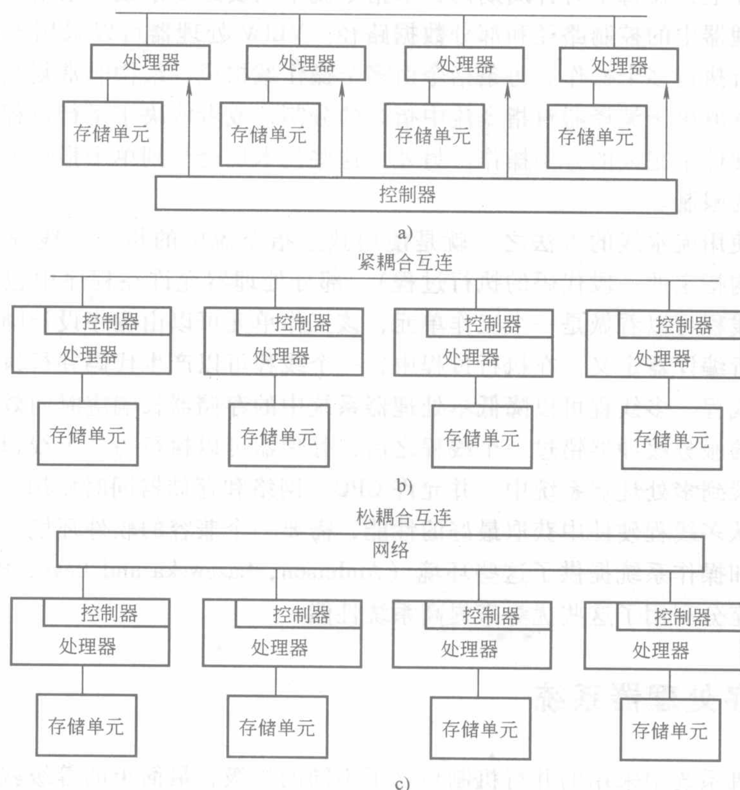


图16-6 系统级并行机制
a）SIMD b）MIMD c）多计算机系统

存储器访问或 I/O 访问)；或者说一旦当前处理过程结束，很难快速分配新的处理过程。MIMD 结构的主要问题是处理器可能会由于不合理的负载平衡而处于空闲状态。如果要实现操作系统（OS），最重要的就是保持资源的高利用率；同时操作系统可以执行系统任务，而且不会产生负载不均衡。

下一代系统是指分布式系统或"多计算机系统"，由于其简单的连通性而大受欢迎。图 16-6c 给出了与处理器无关的网络连接示例。在该网络连接中，每个处理器都是一个独立的实体，并运行独立的操作系统。一个多处理计算机系统通常采用消息机制来在各处理器之间交换数据或指令流。多处理计算机系统的主要问题是消息传递的响应时间以及与分布式存储系统之间的算法映射问题。

16.5　存储器分级体系

高性能的计算机系统都采用分级的存储器体系来提升性能，该体系包含了从小容量、快速 Cache 存储器至大容量、慢速主存的各种存储器。并行机制可以通过存储器分级体系应用到系统中，如图 16-7 所示。Cache 是容量最小、速度最快的缓冲器，通常用来临时存储系统数据，目前广泛应用在处理器中。Cache 存储器充分利用了程序的时间和空间特性。时间特性是指程序中某一项在某一时刻被应用之后，很快就可能再次被应用；而空间特性是指如果程序中某一项被涉及到，邻近的项目也可能很快被涉及到。



图 16-7　将并行机制合并进系统的常用方法

Cache 错误会导致处理器重新从速度较慢的主存中提取所需的数据。平均访问时间取决于 Cache 错误率、从 Cache 中提取数据所需的时钟周期数以及当发生 Cache 错误时从主存提取数据所需的时钟周期数。

Cache 可以划分为独立指令 Cache 和数据 Cache 或标准 Cache，标准 Cache 中包含了指令和数据。独立指令 Cache 通常与 Harvard 类型的体系结构有关，在 Harvard 类型的体系结构中，存在很多独立的指令和数据存储器端口。

在大多数指令提取方式中，单芯片处理器中包含了一个简单的片上指令 Cache。一旦发生 Cache 错误，系统就会向外部存储器发出一个指令请求，这种片下的请求会与外部存储器的数据参考访问产生竞争。

存储器的响应时间是高性能体系结构中的关键。片上和片下访问响应时间的不一致直接促使了下面描述的方案（Hwang，1993）的诞生，即以下 4 种响应时间补充方式：

1）利用预提取指令技术；

2）采用相同的Cache；

3）采用松弛存储器连贯模型；

4）在处理器中采用多上下文或多线程技术。

为了使多指令机制发挥最大的效能，存储器访问时间必须很低，而且必须拓展Cache的带宽。不过，对Cache的访问通常会成为流水线操作中的一个关键时间路径问题。

在多线程处理器中，扩展Cache会对Cache的设计带来一些新的问题。在单线程Cache中，当一个上下文切换时（即处理器开始执行另一个线程），Cache会被冲洗掉，然后从新的线程中提取指令。一个多线程Cache至少支持两种不同类型的指令流，其中不同的线程采用不同的阵列。另一种策略是允许各线程共享Cache，这使得在上下文切换时不必将Cache中的内容冲掉。旧线程的状态保存在处理器中，而且处理器对于各个线程都保存了不同的寄存器文件，而Cache将线程序号和指令发送到处理器的指令缓冲器中。为了给执行单元提供指令，处理器要求在每个时钟周期内都必须提取多个指令到处理器的指令缓冲器中。在这种高效的指令提取方式中，指令Cache访问可能会与Cache线不一致。为了实现Cache线交叉，队列网络可以用来选择合适的指令，并将这些指令提取到指令缓冲器中。通过消除数据总线的复用需求、减小关键定时路径上的I/O时延并提供对Cache的同时读写，可以提升系统的性能。尽管这些操作看上去很简单，但是在设计多线程Cache时必须充分考虑很多因素，例如地址译码和映射、变量指令和数据提取、Cache线交叉、请求协议和队列、流水线和时钟替换策略以及多路访问端口。

通过在各个处理器中设置多个Cache（Cache中包含了主存的一个子集），很多处理器就可以使用相同的存储器数据。当一个处理器对其Cache进行修正时，系统中的其他具有相同数据的Cache也会受到影响。这就需要一个监控机制来刷新Cache中的数据，或使Cache中的数据无效。这些机制通常会在系统中添加相应的硬件，并严格限制与总线型网络互连的实现，这样Cache的刷新就可以得到监控。

16.6 实现过程中的注意事项

在处理器结构的实现过程中，第一步不是构建一个系统来测试电路训练板模型的性能，而是利用近似处理器结构的仿真模型来仿真处理器的性能。其中，对硬件的建模可以分为两种等级。

高级分级模型提供了处理器中不同类型存储器结构的有效分析。模型中的数据流和有效参数可以随不同结构变化而变化。特定结构的低等级仿真（寄存器

转移等级，RTL）通常更加详细地描述了关于实际硬件的性能，这些结构中可能包括 Cache、寄存器文件、变址指令和数据 Cache、存储器控制器以及其他单元。硬件描述语言就是被开发用来描述该等级的电路而设计的。在电路设计建模和仿真完成之后，就可以采用多种技术来实现电路的设计了。

尽管大量的先进技术可以用来提升集成电路（IC）芯片中电路的速度，但是电路速度的提升并不一定会带来计算机性能的提升，因为芯片内和芯片间的时延性能并没有随着电路速度的提升而产生相应的改善。快速逻辑电路之间传输信号所采用的物理方法产生的响应时间限制了计算机性能的提升，不过这种性能的提升可以通过在设计中合并逻辑模块来实现。

这种现象在流水线系统中非常明显。在流水线系统中，最慢的流水线阶段决定了整个系统的时钟速度，这种效应也会出现在硬件电路的各个元器件中，速度的相对提升（即通过提升电路中某一元器件的速度获得的）是由电路中最慢的元器件决定的（如下面的等式所示）。

$$\text{时延}_{\text{电路}} = \text{时延}_{\text{互连}} + \text{时延}_{\text{逻辑模块}}$$

即使通过速度最快逻辑模块的时延接近于零，但是通过互连部分的时延会限制整个系统或逻辑时钟的速度。因此，零逻辑时延系统速度的提升与一定逻辑时延系统的速度提升可以分别描述如下：

$$\text{速度提升} = \frac{\text{时延}_{\text{有逻辑时延}}}{\text{时延}_{\text{无逻辑时延}}}$$

$$\text{速度提升} = \frac{\text{时延}_{\text{互连部分}} + \text{时延}_{\text{逻辑模块}}}{\text{时延}_{\text{互连部分}}}$$

上述这些讨论说明我们必须采用新的方法来补偿物理限制带来的响应时延，而不是通过技术创新来进一步实现速度提升。如图 16-8 所示，芯片间信号传输的时间是高速电路中芯片时钟周期的数倍。芯片间的时延不仅会受到芯片间互连导线的长度影响，也会受到导线之间的互连类型和数量的影响。无论信号是经过一个转接电路还是一个引脚传输到另一种类型的材料或封装中，阻抗和外形尺寸上的不匹配都会产生信号反射，该反射会影响到路径上的传输时延。

实现同步逻辑模块的方式之一是采用自动定时电路或三进制逻辑电路来实现信号的传输。三进制逻辑电路必须配合使用某些光学互连方法；同时，还需要一个逻辑模块来将多值信号转换成数字信号，除非整个系统都是设计在三进制逻辑电路中。另外，整个三进制逻辑电路系统的设计开销也是很大的（逻辑上占据整个系统开销的 50%）。

将系统设计成流水线结构可以使一条指令在芯片间传输时，另一条指令正处于执行的阶段。这种交叠技术可以补偿电路中的响应时延。流水线技术（见图 16-9）不仅可以用来补偿互连时延，也可以通过在芯片的输入端和输出端插入

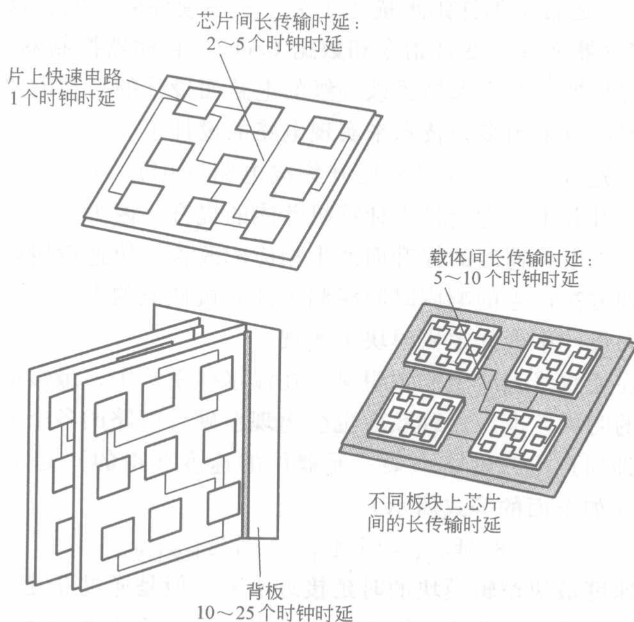


图 16-8 不同等级封装的相对传输时延

注：通过互连单元的时延比芯片时钟周期长

寄存器或锁存器来监听芯片的输入端和输出端（当扫描路径被设计在锁存器中时）。但是，后者会增加流水线的长度和逻辑模块的数量，从而增加系统的响应时延和设计中的电路排版面积。使用流水线技术的主要问题是时钟同步和保持传输时延在每个流水线阶段的对称性。每个寄存器都可以工作在不同的时钟信号上，而各个时钟信号都和其他时钟信号有不同的偏移；也就是说，一个信号到达流水线寄存器的时机和其他信号到达流水线寄存器的时机存在一定的偏移量。时钟偏移量会导致载入流水线下一阶段的信号值出现错误。在理想的流水线设计中，每个流水线阶段对于传输时延都必须是对称的。任何不对称都会导致更长的时钟周期来补偿前面提到的最差情况下的传输时延。理想的流水线系统必须允许时钟倾斜和流水线各阶段的传输时延差存在。

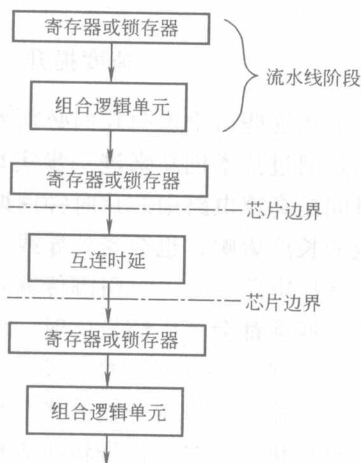


图 16-9 采用流水线技术来补偿互连时延

1. 封装注意事项

确保计算机系统中元器件之间互连的高带宽是非常重要的，而且有些人（Keyes，1991）认为计算机系统的发展是互连技术发展的主要推动力。在早期的关于引脚问题的经验研究中，我们发现在很多设计中，逻辑模块（逻辑模块可以在某个设计分区中实现）的数量和 I/O 引脚的数量之间存在一定的关系，如下所示：

$$I = AB^r$$

式中，I 是指 I/O 引脚的数量；B 是指逻辑模块的数量；A 是指模块的平均大小；r 为 Rent 指数。

这个关系式由于“Rent 规则”而变得非常知名，而 Rent 规则最初来自于 IBM 公司 E. F. Rent 未出版的著作。尽管人们做了大量的研究来反驳 Rent 规则，或者找到了在大多数情况下可以克服引脚限制的体系结构，但是与给定引脚数相关的逻辑模块数量还是存在一定的限制。上面的关系式在很大程度上影响到了某些高速技术的设计；在这些高速技术中，引脚占用了大量的空间，也消耗的了大量的能量。根据封装约束条件来减少引脚的数量，就意味着对于大多数设计来说，逻辑模块必须很简单，也必须很少。

RISC 设计可以提供较少的引脚数量和较小的布线密度，目前已经广泛应用于 VLSI 设计中。在详细的体系结构中，我们已经观察到引脚数量和逻辑模块数量之间的某种关系。由于每个互连部分中可用数字逻辑模块的复杂度是由经验关系式决定的（例如 Rent 规则），因此降低逻辑模块（芯片）的复杂度通常可以降低互连部分（引脚）的数量。由于实现起来更加简单，因此 RISC 芯片比复杂设计需要的布线密度更小，需要的引脚也更少。

封装技术可以为 IC 芯片提供更多的引脚。大多数封装技术都是为硅技术而开发的，不过，现在这些封装技术也开始向由其他技术构成的系统中扩展。大多数封装技术主要是通过在整个芯片上排版来提供更密集的引脚，而不会将排版区域限制在芯片边界上。有时，这些技术中就包括了倒装晶片技术，该技术将芯片的逻辑面反贴在电路板上，而散热片（用来将热量从芯片上驱散）就成了检测逻辑模块的关键。

2. 技术注意事项

在过去的几十年里，硅半导体技术已经成为了计算机系统中最常见的技术。硅芯片的类型也随着逻辑系列的不同而不同，例如 N 沟道金属氧化物半导体（NMOS）、互补金属氧化物半导体（CMOS）、晶体管-晶体管逻辑（TTL）电路和发射极耦合逻辑（ECL）电路。电路集成度和时钟速率仍然是推动这些硅逻辑系列发展的主要因素，这些逻辑系列为计算机设计师提供了很好的性能保证，而无须应付全新的设计策略。高速技术主要应用于高端计算机系统中。那些可能引

导未来计算机系统发展的技术包括：光学元器件、GaAs 器件、超导体和量子效应器件。

很多高速电路技术都是基于 III-V 材料来开发的，而 GaAs 就是当前大多数器件中最常见的材料（有时也和 AlGaAs 层合并使用）。Long 和 Butner（1990）提出了合理利用 GaAs 电路优势和性能的方案，这些优势和性能包括：更高的电子迁移率（更高的转换速度）、半绝缘衬底以及 GaAs 半导体在光学系统中的主要优势。GaAs 电路正在逐步提升电路的集成度，达到了 VLSI 的水平，而且正逐渐成为商业上进行大规模生产的有效技术。目前，越来越多的设计师将 GaAs 芯片合并到了速度敏感的设计领域中。

GaAs 电路包含很多系列，有些系列只能用于耗尽型场效应晶体管（FET）中，而其他类型的系列可以同时用于增强型和耗尽型 FET（Enhancement and Depletion FET，E/D FET）中。VLSI 基于直接耦合 FET 逻辑电路（Direct Coupled FET Logic，DCFL）和源极耦合 FET 逻辑电路（Source Coupled FET Logic，SCFT）。如果采用增强型 FET 和耗尽型 FET，而不用配合额外的电路，那么 DCFT 电路是目前速度最快的、最简单经济的有效 GaAs 逻辑电路。SCFL 电路（也是采用 E/D FET）具有更强的驱动能力，可以提供很快的速度，但它比 DCFL 消耗更多的功耗和空间。

目前，业界正在研究利用光子技术来实现超快速互连和逻辑电路。光子比电子具有更大的潜在带宽，这一点使光学互连技术具有很强的吸引力。与电子不同的是，光子不带电荷，因此不容易受到传输介质的影响。事实上，即使是在高集成度的光学电路或信号中，也不存在串话或电磁场效应。尽管光子束可以传输高质量的信号，但是其逻辑交换过程是很难实现的。也正是由于在信号传输上采用光学技术可以避免各种干扰，这同时也使得光学技术的逻辑交换很难实现。目前，大多数光学元器件或系统在切换到更高速度的过程中具有很高的响应时延，但是，光学互连技术可以为芯片之间或处理器之间的通信提供更宽的带宽。光学互连技术的时延与信号的光电转换过程有关，该过程会增加传输时延。

在导线连接中，传输特性限制了流水线各个阶段导线的使用。反射回来的电信号必须在新的信号进入线路之前得到充分的抑制或衰减，这使得线路上某一时段内只能有一个数字信号在传输。导线互连只能作为流水线的一个简单阶段，如图 16-9 所示。而光学互连由于其传输特性可以处于流水线更高的阶段。一旦一个信号进入光学互连部分，其他新的信号也可以在规定的时间间隔进入线路中进行传输，而不会产生信号干扰。因此，光学互连可以代表流水线设计中的很多阶段，如图 16-10 所示。

目前处于开发阶段的两种类型的光学互连技术分别是：纤维光学和自由空间。纤维光学系统可以在与导线互连相同的速度条件下提供更高的带宽互连。与

传统的导线和通道互连电路相比，纤维光学系统中更快的交换信号可以传输更远的距离。由于任何两个位置之间传输时延的基本物理限制是光的速度，因此自由空间光学系统就成了互连技术的最终选择。光学互连的主要缺点是目前可用的光学技术在单位面积上的集成度没有传统互连技术的高。

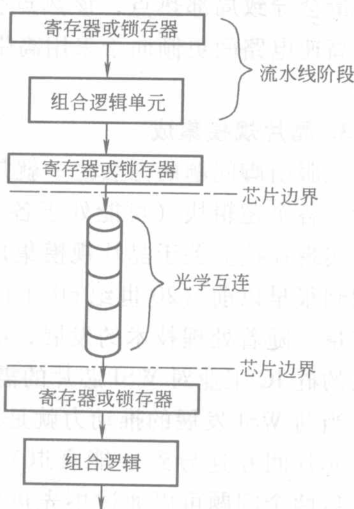


图 16-10 在流水线系统中采用光学互连技术

在纤维光学系统中，从源头到目的地都采用电缆进行传输。电缆通常很宽（常见的是中心距离为 250μm 的电缆），而且很难连接，存在振动问题。光导材料必须通过后续的载体处理过程（包括导线制造和芯片附件焊接程序）来维持低损耗，而且使用的光导材料必须与载体的温度扩展系数相一致，并且涵盖 IC 的工作温度范围。采用纤维光学电缆的好处是在电缆中可以通过器件（如定向耦合器）实现简单的交换功能。

在自由空间系统中，只需要一个检波器和一个光源，就可以将信号从电路板上传输到另一个自由空间。在自由空间中，光纤连接到电路板上时存在的问题避免了，但是光束传输到下一个邻近电路板的传输距离受到了限制。检波器和发送器的布置成为了实现更小特征尺寸的关键，而更小的特征尺寸是提供高密度信号传输所必需的。对于自由空间和纤维光学系统来说，较低的集成度或位置限制使得光学互连技术在某些计算机体系结构中失去了吸引力。

光学互连技术的近期应用主要集中在处理器之间的通信，而光学存储将成为下一个引领光学应用发展的领域。目前，转换电路在进行信号光电转换时还存在严重的传输时延。在光学互连系统中存在两种先进的材料，分别为 GaAs 和 InGaAsP/InGaAs/InP。GaAs 激光系统是目前发展最快、应用最成熟的技术；而 InGaAsP/InGaAs/InP 光电子集成电路（OptoElectronic Integrated Circuits，OEIC）是下一个主要竞争技术，将会与当前可以提供长距离激光功能的技术（例如城域网络和 CATV 分布的纤维光学连接）产生竞争。

降低芯片和互连部分的温度可以提高 FET 电路交换的速度，而且可以降低互连阻抗。互连阻抗的降低可以使互连部分的衰减更小，从而得到更高的带宽。在低温条件下，这些效应既可能在常用的导体中出现，也可以在超导体中出现。由于低温系统的维持需要额外的功耗，因此室温操作通常更受欢迎。芯片上产生

的热量会导致局部热点，该热点会影响其他电路的运行。随着高温超导体的发展，高速电路间更倾向于采用高带宽超导体来实现互连，以获得更快的系统时钟速率。

3. 晶片规模集成

克服引脚问题的方法之一就是通过在一个单晶片上制造芯片组来减少引脚的数量。各个逻辑块（可能处于各独立的芯片上）之间通过晶片上的高速金属互连来实现连接。关于晶片规模集成（Wasfer Scale Integration，WSI）的研究可以追溯到很早以前（20 世纪 60 年代），那个时候的工艺可以提供一定的小型硅芯片产量。随着处理技术的发展，电路集成度上升到了一定的水平，但同时忽略了新兴的硅 IC 工业对 WSI 芯片的需求。

当前 WSI 发展的推动力就是对逻辑模块之间提供高带宽通信的需求（不会由于芯片间互连导致性能衰退）。WSI 的问题主要是产量不够、散热和引脚问题。后两个问题可以通过更先进的封装技术（如倒装晶片和水冷式封装）来解决。如何在晶片的批量生产中获得足够的产量可能成为了 WSI 系统应用的主要障碍。对缺陷的处理会清除掉晶片上的大部分逻辑电路，因此，大部分 WSI 方法采用复制的或多余的逻辑模块来代替有缺陷的逻辑模块，这些复制的逻辑模块可以重新进行静态或动态配置。这样，如果有足够的备件来代替有缺陷的逻辑模块，那么就可以构建一个正常运转的晶片了。

与 WSI 相匹配的体系结构非常简单，而且结构容易复制。随机存取存储器（RAM）芯片设计就是 WSI 最成功的应用。这是因为 GaAs 正处于发展阶段，而 Si 已经在 20 世纪 60 年代早期就发展成熟了，因此 WSI 就为处理器的规模生产提供了另一种手段。由于 GaAs 的脆弱性，大型晶片是无法实现的。采用附着在其他材料上的无缺陷 GaAs 芯片的封装技术可能比 WSI 更适合某些设计，主要原因是 GaAs 的脆弱性、GaAs 晶片的限制尺寸以及 GaAs 材料的成本。

在混合 WSI 中，无缺陷芯片附着在衬底上，并通过常见的处理技术来实现互连（注意：混合也同样适用于微波元器件。混合 WSI 元器件也称为“晶片传输模块（WTM）”）。这种伪 WSI 技术可以提供一定的片上连接密度和性能，而且不必通过维修机制来提高产量；其互连单元具有和 IC 处理中相同的特征尺寸和缺陷等级。混合 WSI 元器件中只有很少几个面可以用来发送信号，并保持较低的互连密度。混合 WSI 已经成为 GaAs 芯片实现的一种方式，在该方式中光电子元器件被集成到高逻辑密度的 Si 芯片中。

4. MCM

MCM 具有高性能的互连技术，可以实现高性能的引脚和高制造水平的元件。MCM 在结构上与混合 WSI 非常接近。其中，有源冲模被直接焊接一个衬底上，主要用作芯片间的互连线，并在单冲模和印制电路板（PCB）之间实现一个高水

平封装。MCM 和混合 WSI 之间的主要区别是衬底和互连的尺寸。MCM 的衬底可以和晶片一样大，但是通常情况下比晶片小几个配线薄膜层大小（可能为 35 个或更多）。

MCM 互连单元的特征尺寸大约是片上互连单元的 10 倍，这使得片上互连单元的大小可以几乎降为零，该互连单元在处理缺陷时失去了效能。为了确保只有性能良好的冲模被焊接到衬底上（通过在一定温度和电压条件下将冲模烧烙到衬底上），MCM 的产量可以处于一个可接受的水平。MCM 的基本结构如图 16-11 所示。

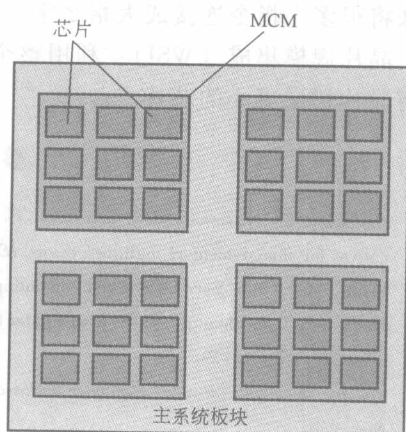


图 16-11　MCM 的基本结构

名词解释

算法：用来解决问题的一组明确定义的步骤或程序。

体系结构：计算机的物理结构，主要是指其内部元器件（寄存器、存储器、指令集、输入输出结构等等）及其相互作用的方式。

复杂指令集计算机（CISC）：一种计算机设计样式，在该样式中指令类型和编址模式存在很少限制。

上下文：执行处理过程所需的信息，包含了程序计数器、地址和数据寄存器以及堆栈指针的内容。

每条指令占用的时钟周期（Cycles Per Instruction，CPI）：一种性能测量方法，用来评定某个详细设计的效率。

多线程：同时执行多条指令流或线程。

流水线：将指令的执行过程分割成重叠的步骤，这些步骤可以和其他指令同时执行，每个步骤代表一个不同的流水线阶段。

精简指令集计算机（RISC）：一种计算机设计样式，在该样式中只执行最常用的操作指令，寻址模式被限制在寄存器中，特殊负载/保存模式被限制在存储器中。

超流水线：一种流水线，其深度被拓展，以允许更多的交叠指令执行。

超标量体系结构：每个时钟周期内可以动态处理多条指令的一种硬件性能。

三进制逻辑：具有 3 个有效电压等级的数字逻辑。

超长指令字（Very Long Instruction Word，VLIW）：一种编译器技术，可以

用来将很多小指令连接成大指令字。

晶片规模集成（WSI）：利用整个半导体晶片来实现一种设计，该设计中无须将晶片切成更小的芯片。

参考文献

[1] Anderson, T., Lazowska, E., and Levy, H. 1989. The performance implication of thread management alternatives for shared-memory multilprocessors. *IEEE Trans. On Computers* 38 (12): 1631-1644.

[2] Flynn, M. J. 1966. Very high-speed computing systems. *Proc. Of the IEEE* 54 (12): 1901-1909.

[3] Hennessy, J. and Jouppi, N. 1991. Computer technology and architecture: An evolving interaction. *IEEE Computer* 24 (9): 18-29.

[4] Hwang, K. 1993. *Advanced Computer Architecture: Parallelism, Scalability, Programmability.* McGraw-Hill, New Yorw.

[5] Keyes, R. W. 1991. The power of connections. *IEEE Circuits and Devices* 7 (3): 32-35.

[6] Long, S. I. and Butner, S. E. 1990. *Gallium Arsenide Digital Integrated Circuit Design.* McGraw-Hill, New Yorw.

[7] Smith, B. 1978. A pipelined, shared resource MIMD computer. *International Conference on parallel Processing*, (Bellaire, MI), pp. 6-8. (Aug).

备注

读者还可以参考以下书籍、期刊或会议资料：

[1] Baron R. J. and Higbee, L. 1992. *Computer Architecture.* Addison-Wesley, Reading, MA.

[2] Mano, M. M. 1993. *Computer System Architecture.* Prentice-Hall, Englewood Cliffs, NJ.

[3] Patterson, D. A. and Hennessy, J. L. 1991. *Computer Organization and Design: The Hardware/Software Interface.* Morgan-Kaufmann, San Mateo, CA.

会议学报：

[1] “国际并行处理技术会议（International Conference on Parallel Processing）”。

[2] “国际计算机体系结构座谈会（International Symposium on Computer Architecture）”。

[3] “国际 VLSI 设计会议（International Conference on VLSI Design）”。

[4] “超级计算会议（Supercomputing）”。

杂志和期刊：

[1] *ACM Computer Architecture News*。

[2] *Computer Design*。

[3] *IEEE Computer*。

[4] *IEEE Micro*。

[5] *IEEE Transactions on Computers*。

第17章 软件设计与开发

Margaret H. Hamilton

17.1 引言

一个计算机系统可以巧妙地被比喻成一个生物实体，我们称之为“超有机体”。硅超有机体就是由软件、硬件、人力实体以及它们之间的互连关系（例如互联网）构成的一个有机体；在该有机体中，所有的组成部分都占有一个重要的地位。硅超有机体本身也是更大的有机体的一部分，例如业务系统。这种业务系统可以是一个医学系统，其中包含了病人、药物、医药公司、医生、医院和健康护理中心；也可以是一项太空任务中，其中包含了太空飞船、宇宙法、任务控制以及宇航员；或者是一个研究基因的系统，其中包含了资助单位、资金、研究人员、研究项目以及基因；或者是一个金融系统，其中包含了投资者、货币、政策、金融机构、股票市场以及世界经济的繁荣状况。

无论业务系统是政府的、学校的，还是商业的，计算机系统就像是一个生物实体一样，必须随时适应不断变化的各种要求。类似于其他有机体，业务系统既有物理的基础设施，也有运作上的各项政策，这些政策指导着企业的发展方向和发展速度，但有时也束缚了企业的发展。

与生物超有机体不同，软件可以随时进行修正，而生物超有机体必须历经很多年才能实现很微小的遗传性进化，因此软件在发展适应性方面比生物实体先进。业务系统规则和基础设施的持续性在“软件可以变化多快”和“系统可以多快接受软件的变化”之间建立了一个自然的联系。

作为硅超有机体的“头脑”，软件控制着整个实体的运转。但是有一点必须牢记，软件仍然是人类创造的。

在本章中，我们将讨论软件的基本原理（即软件是什么以及软件是如何发展的）和软件工程的基本规则，该基本规则是一种方法论，通过这种方法论我们可以将各种想法转移到软件中去。

17.2 软件的概念

无论在支持商业功能或对物理器件的控制方面，软件都是逻辑处理过程的具

体实现方式。正是由于软件的实质就是作为处理过程的具体体现方式，因此其应用可以非常广泛（如进行复杂机构的建模），也可以非常狭窄（如实现离散数字运算）。在前者中，软件在“二次工程”业务系统之间存在很多相互联系，这种联系可以加快业务系统的发展速度，或者说加快操作模型的建模速度，该模型建立之后就被转换成了应用套件；最后，软件就专注于实现各种算法了。因此，软件具有非常广泛的潜在应用范围，只要设计合理，软件就可以长期使用。

但是，也有人将软件只定义为利用编程语言描述编辑过程时产生的代码。而更广义、更准确地定义中包含了要求、规范、设计、程序项目、文档资料、流程、规则、测量方法和数据，以及用来生成、测试、优化并重复使用的工具和软件的实现过程。

软件出现多个定义是对软件开发过程自身认识产生混淆的直接结果。1991年，软件工程协会（Software Engineering Institute，SEI）的研究说明了这个令人吃惊的问题。SEI提出了一种对机构的软件过程成熟度进行分类的方法论，该方法论将软件过程成熟度分为5个等级，第1级为初级（在该级别中，软件还没成形），第5级为优化级，该级别中所有的方法、流程和机制都集中在了如何进一步增强软件的可靠性上。该研究表明，在美国86%的受调查组织机构都处于第1个等级，即符合我们常用的形容词“临时的”、“依赖于英雄式人物的”和“混乱无章的”。即使在当前基于互联网的各种复杂应用中，我们可以惊奇地发现处于第1级的组织机构的比例仍然在不断上升。导致企业混乱的罪魁祸首就是企业自身，这些企业认为所谓的秩序是由某些技术服务来控制的，这种想法只会使企业变得更加混乱和复杂，或者他们至少必须为他们的想法付出更大的代价。

要想摆脱企业的这种混乱不堪的局面，为企业打造良好的企业秩序，必须深刻理解软件的构成及其开发过程。根据自然科学世界的观点，生命原理是指当你将很多简单的事物放在一起时，生命就是其中涌现出来的复杂事物。例如，水分子就是十分活跃的，当你将一大堆这种水分子倒进玻璃杯时，在水的表面就会产生一圈一圈的波纹。如果你将足够多的这种分子合并在一起，你就可以“卷起”一片大海了。同样的道理，软件自身只是一串代码，是一件非常简单的事物；但是将足够的代码合并在一起，你就可以创造一个复杂的程序了。如果再添加一些程序，你就可以创造一个可以“将人类送上月球”的系统了。

尽管系统整体确实比它组成部分的总和要大，但如果要想系统整体运转和控制正常，我们必须彻底掌握各个部分。类似于物理实体，软件也可能会产生“磨损”而需要维护，或在基础系统中必须适应用户需求的变化而不断进行更新。熵是软件中非常重要的一项属性，尤其是对于第1级的组织机构。

处于最低编程级别的软件称为“源代码”，源代码不同于可执行代码（可以

被硬件直接执行来实现一个或多个指定功能的代码)，源代码是采用一种或多种编程语言编写的，而且自身无法被硬件执行，除非被转换成机器语言代码。编程语言是指一组文字、字母、数字和缩写的记忆方法（函数)，由一定的语法进行组织，用来向计算机描述一个程序（由一串串源代码构成)。编程语言存在很多种类，其中很多编程语言都是为某些特定应用度身定做的。C语言就是目前应用最广泛的编程语言，可以应用在工程和商业环境；而面向对象语言（如C++，Stroustrup，1997）和Java（Gosling等，1996）也同样适用于这些环境。事实上，Java现在已经成为了互联网基础应用的首选编程语言。在前些年，工程应用中通常使用的编程语言包括：Fortran、Ada（政府应用）和HAL（太空应用）（Lickly，1974)，而在商业应用中通常青睐于面向商业的通用语言（COBOL)。很明显，我们可以发现在任何组织机构中，没有指定使用哪一种语言。不过，我们可能会希望有更多种类的语言被开发出来。

无论是C语言（Harbison，1997)、Java、COBOL、C++，还是C#或者其他语言，编程语言都提供了对逻辑结构进行编码的能力，在编码过程中，必须注意以下事项：

1）用户接口：提供了一种机制，通过该机制终端用户可以输入、观察、处理和访问组织机构计算机系统中的信息。研究表明，在提供了可视化用户接口之后，企业的生产力会大幅上升。图形用户界面（Graphical User Interfaces，GUI）就是各个操作系统用来显示自身变化的一种机制。常见的图形标准包括UNIX系统（包括Linux）的Motif、Macintosh操作系统的Aqua以及PC系统的Microsoft Windows。

2）模型计算：执行计算过程或程序设计中的各种算法（解决问题时采取按部就班的流程)。例如，过程控制、工资计算或卡尔曼（Kalman）滤波器。

3）程序控制：通过比较、子程序、调用其他程序以及反复执行程序的逻辑功能来实现的控制功能。

4）消息处理：消息处理有很多种类，帮助消息处理就是程序中用来帮助终端用户的结构。错误消息处理就是在输入、输出、计算、报告、通信等过程中，程序用来通报错误并从错误中恢复的自动功能。在面向对象的开发环境中，消息处理是指程序对象将信息传递给其他程序对象的能力。

5）数据移动：程序将数据保存在数据结构中。在程序中，数据可以在数据结构之间移动；也可以从外部数据库或文件中转移到内部数据结构中，或者从用户输入端转移到程序的内部数据结构中。类似地，数据也可以从内部数据结构转移到数据库中或者终端用户的用户接口中。数据移动包括数据分类、搜索和格式化，以及为下一步操作准备数据的相关数据操作。

6）数据库：数据（对象)、主题信息或相关主题信息以及系统信息的集合

体（例如，卡车中的引擎或者组织机构中的人事部）。一个数据库可以包括各种对象（如格式和报告）和系统各项属性（例如，企业中人事部关于员工的各种信息）。数据库的组织原则是必须方便计算机用户的访问，其中的数据是指对属性、概念或计算机用户处理指令的描述。这些属性无法显示、更新、查询和打印，但是可以形成报告。数据库可以以多种方式保存数据，这些方式包括相关形式、分级形式、网络形式或面向对象形式。

7）数据声明：向程序描述数据结构的组织机制。在数据声明中，详细的数据结构必须与其类型联系在一起（例如，关于某个员工的数据可能是人员的类别）。

8）对象：一个人、地点或事件。一个对象中包含了数据和用来处理数据的一组操作。对于人来说，对象就是指他知道什么（称为“属性”）和他可以做什么事（改变自己或与其他对象进行互相作用）。例如，在机器人系统中，一个对象中包含的可能是将金属外壳转移到合适位置的各种功能信息，或者同时与另一个机器人合作转移另一个对象的信息。对象之间可以通过通信媒介（如消息传递、无线电波、互联网）进行相互通信。

9）实时：满足关键时间需求的软件系统。软件的正确性取决于计算的结果，同时还取决于结果的产生时间。实时系统具有各式各样的要求，例如在指定时限内完成某一项任务，并对计算机外另一个处理过程的相关数据进行处理。这一类的应用包括：交易处理、航空电子系统、交互式事务管理、汽车系统和视频游戏，上述这些都是实时系统的典型应用。

10）分布式系统：很多独立的互连处理过程协作形成的系统（例如，多个计算机上的多处理过程）。客户/服务器模型就是当前最常见的分布式形式。在该模型中，客户产生一个分布式功能，而服务器执行该功能。

11）仿真：通过另一个系统对一个实际系统或抽象系统性能中的某些选定特性进行描述。例如，一个软件程序可以对飞机、机构、计算机或另一个软件程序进行仿真。

12）文档资料：包含了对要求、规范和设计的描述，同时还包含了对程序操作过程的描述，该描述保存在程序内部；另外还包含了生成文档，该生成文档用来描述每个程序在更大的系统中是如何操作的。

13）工具：用来设计、开发、测试、分析以及维持另一个计算机程序系统及其文档设计的计算机程序。工具中包含了代码生成器、编译器、编辑器、数据库管理系统（Data Base Management System，DBMS）、GUI 编码程序、仿真器、调试器、操作系统和软件开发以及系统工程工具（部分来自于以前的工具，称为“计算机辅助软件工程（Computer Aided Software Engineering，CASE）”工具）。在 CASE 工具中集合了一组工具，包括上面列出的那些工具。

至此，尽管读者应该理解了一串源代码的动态本质了，但是源代码是否适合软件的“超有机体”，还存在很多变数。这种变数包括读者的行业以及各个机构的软件开发样式。

作为一个基本单元，一串源代码可以与其他源代码联合并形成很多新的东西。在传统软件开发环境中，很多串源代码可以构成一个程序，有时称为“应用程序”或者“简单应用程序”。但是源代码本身是无法执行的，因此，源代码首先必须通过一个编译器来生成目标代码。然后，目标代码通过一个链接器来生成可执行代码。其中，编译器本身也是一段程序，其功能是双重的。编译器首先对源代码进行语法错误检查；然后，如果没有发现错误，就会为指定的操作系统生成目标代码，这些操作系统包括 UNIX、Linux（UNIX 的一种）、Windows 等。操作系统可以看做是一个管理程序，该程序管理着应用程序在其控制之下的运行过程。由于每个操作系统（和计算机体系结构）之间是互不相同的，因此，某一个操作系统编译过的源代码是无法在其他操作系统中执行的，必须重新进行编译。

解决一个复杂的商业或工程问题，通常需要多个程序。利用多个程序的协作来解决问题的过程称为“应用系统”。现在面向对象的开发技术已经将程序的概念进行了重新组合，取而代之的是“对象”的概念（Goldberg and Robson，1983；Meyer and Bobrow，1992；Stefik and Bobrow，1985；Stroustrup，1994）。

程序可以看做是一段关键代码，在解决问题的过程中，用来实现很多功能，而无需考虑对象问题；而对象是与代码相关的属性，代码各种功能在实现过程中只能使用与其相关的那一类对象。相比由传统方法形成的系统，通过合并对象（类似于合并分子）可以形成效率更高的系统；同时，软件的开发也会变得更快，产生的错误更少。由于对象可以重复使用，因此，一旦通过测试和执行，对象就可以放入数据库中，以供其他开发者使用。数据库中的对象越多，开发新系统的速度就越快、越容易。另外，理论上，由于重复使用的对象已经经过实践的检验（例如，通过测试并且没有错误），因此面向对象的系统存在缺陷的几率非常小。

构建程序或对象的过程称为“软件开发”或“软件工程”。这个过程由一系列的步骤或阶段组成，通常称为“开发生命周期”。这些阶段（至少）包括：需求分析阶段（在该阶段对商业或工程问题进行剖析和理解）、规范制定阶段（在该阶段对如何实现需求做出决定，例如软件和硬件分别需要配置哪些功能）、设计阶段（在该阶段设计从 GUI 到数据库、运算法则、输出端的所有内容）、编程阶段或实现阶段（在该阶段利用多个支持人工编码的工具来自动生成代码）、测试或调试阶段（在该阶段对代码进行测试，该测试与商业测试不同；同时对程序中发现的错误进行检错和纠错）、安装阶段（在该阶段系统被安装到产品中）

和维护阶段（在该阶段对系统进行修正）。但是，不同的人开发系统的方式是不一样的，因此，他们对软件工程的概念和观点也不尽相同。

17.3 软件工程的实质

工程师们通常采用“系统工程”这个词来表示一个非软件系统的需求规范、设计和仿真过程，例如桥梁系统工程或电子元件系统工程。尽管软件可用来进行仿真，但是它只是系统工程处理过程中的一部分。另一方面，软件工程只关注软件产品，而不是其他事物。在20世纪70年代，行业评论家们开始注意到大规模系统的生产成本在快速增长，而且很多项目都失败了或者生产出来的产品非常不可靠。软件的发展遇到了前所未有的危机，其中最主要的关注点包括以下几个方面：

1）程序设计师的效率：在20世纪80年代的政府机构中，一个使用C语言的一般开发人员平均每天只能编写出10行代码（而商业机构的一般开发人员每个月也只能编写30行代码）。现在的标准是每天大约2~5行，同时系统的需求远比以前高出多个数量级，因此会造成严重的产品积压。国家软件质量实验（NSQE）在2002年完成一项为期10年的研究项目，该项目对现在的系统和10年前的系统进行了比较。该项目显示，人们开始以为现在的系统在提供了更“先进”的开发工具（商业包装的增值效果和面向对象的语言）条件下，其效率会远高于以前的系统，但是后来他们发现错了。现在的系统效率不仅没有提高，反而有所下降（NSQE，2002）。另外，NSQE还总结出，软件的设计和开发方法必须取得重大突破，才能在缺陷率上实现10%的下降。

程序设计师的效率取决于各种奇特的想法：从专业技能到问题的复杂性都可能直接影响到开发出来的程序（Boehm，1981）。软件工程质量和产量的测量科学称为“测量机制（Metrics）”。在软件工程的各种样式中，有很多种软件测量样式。现在软件工程的各种测量形式是很复杂的，而且必须考虑以下各个方面：成本、产品上市时间、优先项目的效率、数据通信、分布式功能、性能、重复使用的配置、处理速度、在线数据登录、终端用户效率、在线更新、处理过程复杂度、重复使用率、安装容易度、操作容易度以及运行环境的多样性。

2）缺陷消除成本：影响程序和对象调试成本的那些变化因素，这些因素也会直接影响到程序设计师的效率。因此，人们注意到，程序的测试和修正也会耗费大量的人力和物力。

3）开发环境：开发工具和开发习惯会在很大程度上影响到软件的质量和数量。现在，大多数设计和编程环境中只包含了一些开发完整系统所需的内容。其中，生命周期开发环境就是一个很好的例子。这些工具中大多数可以看做是生命

周期的前半部分（例如，用来分析和设计），或者是生命周期的后半部分（例如，用来生成代码）。在市面上，很少有综合性的工具（同时能实现前半部部分和后半部分的功能）；而将仿真、测试和交叉平台综合在一起的工具就更少了。因此，同时将系统设计集成到软件开发过程中就更少见了。

4）GUI 开发：开发图形用户界面（GUI）是一个非常复杂和昂贵的过程，除非采用非常合适的工具。系统从主机开发环境到工作站 PC 应用环境的转移过程可以看做是 GUI 开发程序进入市场的过程。但是，大多数基于 GUI 的工具还不具备开发整体系统的能力（例如，不仅仅是前端用户的处理部分）。因此，GUI 也会开发出不完整的或有错误的系统。为了提高效率，GUI 生成器必须完全集成到软件开发环境中去。

基于上述这些方面，现在大多数系统都需要比原始开发过程更多的可分配资源。事实上，Lientz 和 Swanson 在 1980 年就论证了一个观点，即当时的系统开发问题远比 20 世纪 70 年代开始时认识到的问题多。软件开发实际上是非常复杂的，而且软件开发工程小组还会受到时间和预算资金的限制。另外，Jones 早在 1977 年就已经对软件开发人员进行了详细记载。

实质上，很多软件工程都试图扭转不断下降的产量和质量。但是，这些努力都失败了，因为这些方法只能治标，而不能治本。事实上，软件工程完全取决于对系统软件（如操作系统）和硬件以及软件工程所处的商业环境的理解。

SEI 的过程成熟度图形非常准确地指出了软件开发的根本问题。调查表明，86% 的组织机构仍然停留在非正规或者混乱的水平，而只有 14% 的组织机构构建了正规的软件工程。86% 的机构只是通过编写代码来处理业务问题，即使他们发现了一定的软件工程规律，但很可能那些规律也已经不再符合业务发展的需要了。

在 20 世纪 70 年代，结构方法论非常流行。对于大多数课题来说，尽管存在一定的问题（例如，不同版本的结构技术，包括流行的 Gane/Sarson 方法和 Yourdon 方法），但是结构方法论在批处理时代为各种实用系统的开发提供了方法。在那个时代，即使带有完全无声终端的在线系统也是一个革命性的概念，而图形用户界面更是如同柏林墙的倒塌一样，是一件不可能想象的事情。

尽管时代已经改变，而且如今的硬件比结构技术时代的硬件先进一千倍，但是结构技术仍然保存下来了。同时，尽管这些技术发明者的目光已经转移到了更加先进的技术，但是结构技术和更现代的软件开发和系统工程环境还是进入了市场。

在 1981 年，Finkelstein 和 Martin（Martin，1981）为更多地面向商业用户（例如，那些亟待解决基于数据库问题的用户）普及了信息工程概念。现在，信息工程在 CASE 投资的主流开发商中已经非常受欢迎了。信息工程实质上是结构

方法的重新定义。但是，信息工程不是非常注重结构方法中的数据，而是注重整个机构或系统的信息需求。在此，业务专家定义了高级信息模型和详细的数据模型。最终，系统就是根据这些模型进行设计。

结构工程和信息工程方法论都来源于面向大型机的商业应用。而如今，客户/服务器技术（机构的数据可以在位置分布式的服务器之间传输，而终端用户的GUI可以实现逻辑处理过程）已经淘汰了这两种技术。事实上，很多在商业应用中面临的问题与面向工程环境（如电信和航空电子设备）中面临的问题是不同的。客户/服务器环境的主要特性是多样性。一个组织机构可能会将数据保存在多个数据库中，而程序由多种编程语言构成，而且采用了多个操作系统，因此，其GUI系统也是不同的。由于软件开发的复杂度在这种新的环境中增加了数百倍，因此，就需要一个更好的方法论来支持。如今，面向对象技术解决了这个问题。在给定了客户/服务器的复杂度之后，嵌入在程序中的代码就没有了足够的灵活性来满足这类环境的需求了。我们已经讨论过了，通过反复使用对象而不是大型程序进行编码时，可以实现很好的软件灵活性、效率和质量。但是，面向对象的开发过程也是“一把双刃剑”。

采用面向对象技术可以大幅提高软件的效率，但同时如果使用不当，面向对象的开发过程也会产生比结构技术更多的问题。这个道理很简单：因为起点更高了。面向对象的开发环境比任何其他开发环境都复杂，面向对象技术面临解决的问题也比其他类型的问题更加复杂，而且也没有传统的面向对象方法论和配套工具来帮助开发人员开发真实可靠的系统。面向对象存在很多种类，但这种多样性同时也带来了很多风险。因此，在进行面向对象开发之前，必须充分考虑以下几个开发要点：

1）综合具有很大的挑战性，在开始时就需要充分考虑。在传统系统中，开发者主要依靠各种不匹配的建模方法来获取系统定义的各个方面。无论系统中的综合方式是指对象之间的综合、模型之间的综合、阶段之间的综合，还是应用类型之间的综合，这个过程都是非常困难的；其中，设计和开发过程中关键产品版本的不匹配（关键产品只能用来开发某一种应用）是主要原因。因此，综合只限于各种开发者的器件开发过程。最终实现的系统也可能会变得很难理解，系统的对象也很难描述；而其中最大的缺陷是与实际系统完全不相符；而接口相互之间通常也不协调，从而错误也会贯穿整个开发过程。因此，采用这种方式定义的系统是非常不准确的，而且通常是错误的。

2）错误最小化。传统方法和传统的面向对象方法实际上推动了错误的传播。例如，重复使用含有嵌入和遗传错误的不可靠对象。互联网病毒、蠕虫、数据包风暴、特洛伊木马以及其他恶意的程序都是不可饶恕的，而且必须排除在系统外。从成功的角度来说，错误在开发过程的最开始阶段就应该被消除，包括在

操作系统和其他相关系统中。

3）开发语言应该更加正规。尽管有些语言是正规的，而也有些语言是“友好”的，但是我们很难发现一种语言既是正规的，也是“友好”的。在使用很多非正规语言的开发环境中，缺乏可描述性、过多接口错误是很常见的现象；而在使用正规语言的开发环境中，我们也发现其系统应用在大小和复杂度方面很难与实际系统相一致。最近，很多先进的软件需求语言被开发出来了（例如，统一建模语言，UML（Booch 等，1999）），其中有些语言是直接对几种语言进行了“综合”。但是，更糟糕的是，其中很多都是不需要的，而真正需要的很少。正是由于缺少了正规的部分，因此，就需要一个公共的语言语义学来协调各种语言之间的差别，并消除多余的部分。

4）变化的灵活性和不可预知性的处理必须预先得到解决。在应用系统的开发过程中，我们经常忘记必须将系统的演进考虑在内。在实际应用中，用户会改变其各种想法即软件开发环境和技术。在传统的开发环境中，需求的定义主要集中在用户的应用需要，而没有关注用户需求和环境的潜在变化。转入到一个新环境通常会演变成一个新的体系结构、操作系统、数据库、图形环境、语言或语言配置的开发过程。因此，关键性的功能通常就会被回避，因为害怕对系统不够理解；而且，在开发过程中，对维护的解释不够清晰，而维护恰恰就是系统生命周期中风险最大、成本最高的部分。为了解决这些问题，必须采用各种工具和技术来实现技术交叉和不断变化的技术解决方法，而且必须提供不断变化和演进的体系结构。

5）设计过程中的“综合症”必须得到解决。通常，开发者被迫按照规范技术来进行开发，规范技术中没有开放式的结构，如精确的数据库计划或 GUI。试图为这样的技术找到一个解决方案是一件很难的事；更糟的是将其中一部分技术作为系统的可重复使用技术，而该系统并不依赖于该类技术。这些问题在一个组织机构中通常会通过彻底的理解和正规的商业惯例以及执行过程得到解决。

6）开发者必须为并行机制和分布式环境做好准备。通常，如果不知道系统的目标是实现一个分布式的应用环境，那么开发者首先会定义并开发一个简单的处理器环境，然后再改进或重新开发分布式的环境——对资源的非生产性使用。因此，并行机制和分布式机制在项目开始时就必须考虑。

7）资源分配必须对用户透明。无论系统是否被分配给分布式、异步或同步处理器，以及无论是选择了两个处理器还是 10 个处理器，采用传统方法进行设计时，设计师和开发者必须考虑采用人工的方式来将上面这些细节合并到应用系统中。在系统需要实现什么规范和系统如何实现规范之间没有明显的界限。因此，在设计过程中，就包含了很多实现方式上的细节问题。一旦这样的资源结构变得陈旧，就必须重新进行设计并重新开发应用系统，而这些应用系统中包含了

旧的设计方法。

8）必须加强可靠的可重复使用功能的创造，尤其是内在功能。传统的系统需求定义中缺少了用来帮助发现、产生、使用并确保通用性的工具。建模设计师被迫使用非正规和人工的方法来将系统划分成各个自然的组成部分以便重复使用。这些组成部分并不能直接“综合”在一起，因此，它们也经常产生错误。由于这些部分相互之间不匹配，因此，就不能作为重复使用的功能。在传统的方法论中，冗余处理成为了实现系统的一种方式。即使是采用面向对象的开发方法，开发者通常在面对其开发的器件时也必须明确应用系统是面向对象的系统，这是因为这些方法并不能支持对象中的所有内在功能。

9）将人工操作降到最低程度的自动操作必须替代“人工”自动解决方案。事实上，自动操作本身就是一个内在的可重复使用处理过程。如果一个系统不能重复使用，那么该系统当然也就不能自动操作了。不过，如今大多数开发过程都不需要人工处理了，现在的系统都是根据低智能自动化工具来定义的，并将这些工具作为输入端。事实上，这些自动化工具主要集中在提供人工处理上，而不是完成实际的工作。例如，开发者在得到了系统需求之后，就是通过人工方式将其转换成代码的。一个处理过程一旦经过自动化的重复使用，那么该处理过程就可以人工地反复执行。这样，即使自动操作可以完成实际工作，但也不能完全适用于各个应用领域或者某一应用领域，否则将会产生不完整的代码，如框架代码（即实际代码的概要）；这样产生的代码通常是无效的，或者对体系结构、语言或者语言的版本是无用的。通常，局部自动操作需要与其他性质相反的局部自动操作或人工处理进行“综合”，其中人工处理用来完成自动操作未完成的处理。

10）运行时间性能分析（算法和体系结构之间的关系）必须基于正规的系统定义。传统的系统定义中包含的关于系统运行时间性能方面的信息是不充分的，这些信息包括算法和体系结构之间的关系。在设计过程中，确定何处不用考虑这些关系取决于根据具体的实现过程和相关测试环境对输出值的分析。系统定义必须充分考虑如何将目标系统与目标环境分开。

11）设计的完整性是实现应用系统的第一步。即使执行过程失败了，或者成功之后，我们也无法确定设计过程是否完全正确。螺旋式的开发过程（一种进化方式）与瀑布式的开发过程（每个开发阶段在下一个阶段开始之前完成）刚好相反，可以用来解决前面提到的问题。通常，一个系统的设计是基于短期考虑的，因为关于系统的知识是在不断发展更新的。因此，开发过程就不可避免会出现失败。

如果上述要点得到了落实，那么软件开发的成本和开发时间就会更少；不过，其中时间是最实质的要点。上述要点现在变得更加复杂了，而且当开发者为

各种分布式环境作准备时，这些要点也就更加关键了；其中分布式环境的发展融合了不断增加的互联网优势。

到目前为止，本章已经阐述了软件的来源，并且努力解释了软件如何演变成为真正的自动化系统“大脑”。不过，类似于人的大脑，这种软件“大脑”的结构必须非常精细，这样才能提高效率、保证质量、增强控制性和可重复使用性。

传统的软件工程模式无法从超有机体的角度来看待软件的开发过程（超有机体在本章开始的部分介绍过），我们只能看到软件开发过程是由一些分立的集成部分构成的，因此，我们必须开发出一种高效的方法论，这种方法论可以为软件工程带来很多好处，这一点在几十年前软件诞生之日起就得到了验证。

从这个角度看，软件工程是由一种方法论和一系列的工具组成的，这些工具用来解决身边的各种业务问题。但是，在第一个工具使用之前，软件工程的方法论就必须得到应用，以帮助确定系统的需求问题。那么，对于前面介绍的各种系统要点问题，软件工程方法论是如何解决的呢？如果组织机构要求开发的系统必须运行在各种各样的分布式硬件平台、数据库、编程语言和 GUI 上，而传统的方法论又无法实现时，软件工程的方法论是如何实现这些的呢？另外，软件的开发如何避免出现上述那些问题而且不会造成软件开发过程出现“事后弥补”的后果？

为了解决上述这些软件要点问题，我们有很多选择，这些选择包括：以同样的方式继续下去；在指定领域中添加能支持一般业务又可以提供紧急帮助的工具和技术；引入更加先进但还是传统的工具和技术来代替现有的工具和技术；采用包含最先进工具和技术、并能将软件开发过程正规化的新软件工程方法，同时利用已经开发出来的软件；完全重新开始，并采用能将软件开发过程正规化的新软件工程方法，同时利用最先进的工具和技术。

17.4　一种新的设计方式

人们对软件及其开发过程中的那些问题和挑战的反应是很无奈的，他们不得不承认软件就是这么回事，只能接受这个“既成事实”。但是，这种反应是不能接受的。我们需要的是软件设计方式上的改进，这种改进的方式能指导我们利用正确的技术和时间来构建系统。首要地，改进后的方式是一种预防性的方式，也就是说，它首先能以正确的方式提供软件设计的框架结构。与传统的软件设计和开发方法相关的问题可以通过系统的定义在“事先预防”就得到解决，这种方法主要集中在如何预防开发过程中已经发生过的问题，而不是任由“事后弥补”

现象再次出现；而且当这些问题在最不合适、最不值得的时机出现后，还要及时解决好。

假设这种方式应用到人体系统中，例如在补牙时，我们必须在牙根出现牙槽之前就将牙洞填补好，这只是一种医疗手段，但是可以预防牙槽的出现。如果我们的饮食非常合理，那么不仅会可以预防牙槽，还可以预防牙洞的出现。如果牙洞出现之后又产生牙槽的话，就是最不划算的事情了；而在牙洞出现之后，才想起填补牙洞就是仅次于最不划算的事情了；相反，一开始就采取措施来预防牙洞的出现就显得非常划算了。预防措施只是一个相对的概念。对于给定的系统，无论是人体系统还是软件系统，第一个目标就是在最大程度上尽早预防，因为任何事情在其生命周期过程中都会慢慢变坏。

采用预防性原理，系统在一开始构建时就可以将开发过程中的问题最小化。一个系统如果是根据各种属性来开发的，那么该系统就可以控制其自身的设计和开发过程，这样就可以得到一个可重复使用的系统，该系统还可以促进自动化功能的生成。每个系统定义都可以根据内在的约束条件对其应用和生命周期进行建模，这些约束条件可以保护开发者不丢弃自己的各种思想。

上面这种方法可以贯穿于整个生命周期过程，首先是需求分析，接下来就是功能分析、仿真、规范分析、设计、系统体系结构设计、算法开发、实现、配置管理、测试、维护以及反向工程。这个过程中涉及到的人员包括：终端用户、管理人员、系统工程师、软件工程师和测试工程师。

采用这种方式时，可以使用相同的语言来定义任何系统的任何属性，还可以将某一属性与其他属性综合在一起。最重要的是，这些属性是与实际应用直接相关的。因此，可以使用相同的语言来定义系统需求、系统规范、设计、功能性的具体设计、资源和所有等级和层次的完整资源分配结构，包括：硬件、软件和人力资源。还可以使用相同的语言来定义人员、导弹或银行系统、感知系统的组织结构，以及实时环境或数据库环境。因此，这种方式非常适合各种行业以及学术界和政府部门。

预防性系统背后的哲学原理是根据可靠系统自身要求来定义的。只有可靠系统才可以用来构建功能模块，也只有可靠系统才能作为一种机制来对构建的功能模块进行综合，以形成新的系统；而形成的新系统又可以成为一个可重复使用的系统来构建其他系统。

有效地可重复使用是一种预防性概念，也就是说，重复使用没有错误的系统可以获得理想的功能，并可以避免开发新系统过程中可能出现的错误和高成本。预防性系统可以为我们尽可能快地解决问题，而不用拖到最后一刻。但是，要想得到一个真正可重复使用的系统，在实现过程或维护阶段中，我们不能从习惯性的生命周期终点开始，而应该从生命周期起点处就开始。

预防性系统是生命原理结构的真正实现方式，在这个过程中，软件的组成部分很自然地合并形成一个整体，这个整体远大于所有组成部分的总和。或者，我们可以从小时候的拼装玩具角度来认识结构性系统，我们可以回忆起小时候在拼装玩具时是从不犯任何错误的。实际上，拼装玩具也是一块一块构建起来的，其中每一块都可以永久重复使用，而且是十分完整和方便使用的。

目前，至少有一种实现方式遵循了预防性哲学原理。尽管这种方式不是主流方式，但它仍然得到了研究组织机构和“开路先锋”们的成功应用，而且现在已经适合越来越多的商业应用了。采用这种方式，用户可以设计各种系统，并完整地构建软件（从系统至软件）；而且系统设计师或系统工程师不再需要彻底理解编程语言或操作系统，而且不会产生界面错误，缺陷率也至少降低 10%；用户可以通过内置语言属性来确保语法的正确性，而且系统具有明确的需求、规范、设计（消除了复杂性、混乱和模糊性）；在实现过程之后还可以保证功能的完整性；系统具有完全可描述性和可演进性（应用上、结构上和技术上）以及最大程度上的可重复使用性；用户可以根据系统规范来生成完整的软件解决方案（在生命周期中完全实现自动操作，包括100%为任何类型或任何大小的系统预备代码的产品）；另外，还有一套工具，其中每一种工具都是根据自身定义和自动生成的；因此，系统具有更高的可靠性、更高的效率和更低的风险。

大多数人会说这是不可能的，至少在可预见的未来是不可能的，这是因为在传统的开发环境中确实是不可能的。但是，在系统的主要组成部分中是可能的，因为有非传统的系统设计和软件开发样式以及相关的宇宙系统语言，这些语言来自于阿波罗板上飞行软件的经验式研究，并经过了30年的发展。除了阿波罗和其他实际应用系统的经验之外，这种样式还来源于系统理论、正规方法、正规语言学和面向对象技术。下面，我们将通过实例来描述这种技术及其背景，为读者阐明预防性方式的潜在优势。

17.5　Apollo 板上飞行软件成果：值得学习的课题

经验式研究从1968 开始，其初始目的是从阿波罗板上飞行软件及其开发过程中学习某些东西，以便为未来的阿波罗任务以及后来的太空实验室和航天飞机任务提供帮助。因此，那时就需要一个更好的方法来定义和开发软件系统；因为，当时的方法（类似于现在的传统方法）无法解决紧急性的问题。那个时候，有一股学习经验的驱动力，即如何使未来的系统性能更好，设计师和开发人员应该继续做些什么，因为他们过去以一直做得很好。通过对阿波罗板上飞行软件的研究发现了一种构建非常可靠的软件的方法。

阿波罗板上飞行软件研究取得的成果是多方面的，不仅仅是为太空任务提供了帮助，也为一般的系统提供了指导，其中有些系统很明显需要很多年才能实现。阿波罗软件指出了哪些系统需要实现以及系统实现的时间限制，它实现了它能达到的最高复杂程度，这也导致未来的其他软件项目跟它比较起来显得很不起眼。如果在最坏的条件下产生的问题能够得到解决，那么这种解决方案就可以适应于任何类型的系统了。

从事后的研究来看，阿波罗工程是一个非常成功的例子，从中我们可以学习到系统、软件和可能出现的各种问题的解决方法，以及各种值得关注的事情。由于阿波罗软件的本质原因，各种类型的错误都有可能出现，尤其是由于软件在当时是根据太空任务计划（如硬件、仿真器以及宇航员的培训（论证飞行软件在整个软件、硬件、人力资源系统中占有多大的地位））来开发的，加上之前没有人到过月球，因此，存在很多未知的变数。另外，开发人员是在严密监视下进行开发的，这在今天是一件不可想像的事情。阿波罗工程以及取得的成就都为我们提供了宝贵的信息财富。下面简单列出几点。

在研究阿波罗工程的过程中，我们发现在经过最后的测试、认证和审定（Verification and Validation，V&V）阶段后，接口错误（这些错误主要是由于各种不明确的关系、配合不当、系统冲突、通信困难、缺乏协调、无法综合而导致的）在所有发现的软件错误中所占的比例高达75%（如果采用传统的开发方法，接口错误所占比率会高达90%）。接口错误涉及到系统的最高级至系统的最低级，以及各个具体低级别的各种数据流错误、优先级和定时错误。另外还发现，44%的错误是通过人工手段发现的（这说明还有更多是自动操作领域的），而且60%的错误在早期的任务中也不经意地存在——这些任务已经过去了，尽管在很多飞行过程中没有出现错误（只有他们自己知道）。但事实是，这么多的错误在早期的任务中确实存在，这是非常令人恐惧的，这意味着很多人在每个任务过程中一直是处在危险之中，同时还说明了在可靠性上还有很多工作要做。尽管在阿波罗任务中没有出现软件错误，但是这只是得益于软件小组和他们采用的测试和开发软件方法。

接下来是关于接口错误更详细的分析，这是因为接口错误既是主要的错误，也是很敏感的错误，因此很难发现。系统在实现过程中，只有接口问题得到了解决，完整性问题才能得到解决；而只有完整性问题得到了解决，系统才具有可描述性。每个接口错误都可以根据系统中定义的预防方式来进行归类；然后通过修改定义来确保预防错误再次发生。这项工作为定义一个可以消除所有接口错误的系统提供了理论和方法论基础。

很明显，每一件事情（包括软件）都是一个系统，而且系统的设计要点只有一个，如同软件工程中的一样。很多事情一起形成了这种观点，即“系

统”的概念。在飞行软件的开发过程中，一旦软件工作组实现了某种需求（系统设计专家将需求“扔给”软件开发人员），软件开发人员就必须成为这方面的新专家。这种机制迫使软件专家们变成了系统专家（反之亦然），同时也说明一个系统就是一个系统，无论是以更高形式的规则出现还是通过软件实现的规则出现。

每个想有所收获的人都会期待那些无法预见的事情出现，这种现象在很多设计中都反映出来了。也就是说，他们学会了按照错误检测和恢复的方式进行思考、计划、设计和实时重新配置，并且经常对系统进行备份。对确定分配惟一处理优先级重要性的认识是阿波罗哲学原理中的一个重要部分，这个目标的实现过程也称为“失败之后再重新开始”，这种关于错误检测和恢复的重新开始方法远比相反的“从哪里失败就从哪里再来”方法先进很多，它不仅简化了软件的操作，也简化了软件的开发和测试过程。系统设计对软件某一部分的主要影响也会转移到软件的另一部分上，这一点说明系统中包含的所有东西都是系统的一部分——即“过去的事情还会重来”综合症。例如，飞行软件选择的异步执行方式（在异步方式中，更高优先级的处理过程会中断低级别的处理过程）不仅允许在实际飞行中有更多的灵活性，而且还提供了更多的灵活性，以便在飞行软件开发过程中实现更安全、更标准的飞行。

很明显，阿波罗工程的测试工作永远不会结束（即使比其他种类的应用系统包含更多的测试等级，而且独立的论证和审定组织还会添加其他的测试方法），这种机制可以促使我们寻找更好的测试方法。另外，传统的方法已经无法解决任何问题了。

随着大多数系统设计和软件开发过程实现了机械化（相同的处理也可以通过开发过程中的每个阶段实现），我们很清楚，这个过程也可以实现自动化。这说明自动开发环境可以实现现在传统系统中很多仍然通过人工实现的功能。另外，对于太空任务来说，软件的开发是突然性的，而且同时要为很多任务的各个阶段进行开发，因此，这也为我们如何成功地在分布式环境中实现软件开发提供了学习机会。

在太空飞行软件的演进过程中，我们可以发现，传统的方法在很多领域已经无法为开发者提供支持了，而且在其他领域中也没有提供很多的自由空间（例如，产生既敏感又严重的错误的可能），同时在某些领域中（例如，要求一个开放式的体系结构并能重新进行实时配置以获得成功的领域）也没有提供足够的自由空间。这样，我们很快就能明白，这种情况下需要一种新的语言来定义系统，该系统的某些“内置”属性在传统语言中是无效的，例如内在的可靠性、综合性、重复使用性以及语言自身的应用才能提供的开放式体系结构能力。另外，开发一套新的工具也可以用来支持这样的语言，同时还可以避免繁琐的机械

化过程而实现自动化过程；这样，就可以避免产生更多的错误并大幅减少测试过程之后的各种需求。从上面的分析中我们可以知道，具有“内置可靠性”的系统可以提高开发效率，从而实现“内置效率”。

对上面这种现象的研究今天仍在继续。其中的关键点就是指系统实质上是一个异步系统，这一点在定义系统的语言中就可以反映出来。系统应该是基于事件驱动的，而且每一项功能应该具有惟一的优先级。实时事件和优先级驱动性能必须成为我们在系统语言中定义系统时采用的方式中的一部分，而不能局限于某一事件，而且在不同的编程语言中必须采用各自指定的数据类型。

语言的目的不是为了驱动某个机器，而是为了很自然地描述各个对象以及各个对象执行的操作。对象本身是分布式的，它们之间的相关作用是实时的，而且是由事件驱动的。这就是说，我们可以定义一个系统，而且系统自身的定义中必须包含实时执行过程中可以用来描述系统自然性能的各种必要功能。应用系统开发者不再需要明确地为即将发生的事件定义进度表了，因为当对象之间相互作用时，事件可能就不会发生了，而事件进度表的定义就是用来描述对象之间的相互作用的。

上述这些表明，一个宇宙系统语言可以用来处理一个系统各个方面的事情，包括：计划、开发过程、应用以及演进；这就意味着，所有在系统生命周期内一起工作的系统都可采用相同的语义学来定义。

17.6 “事先预防”式开发

对阿波罗成果的分析结束之后，下一步就是根据阿波罗的“思想”来建立一个数学模式，该模式是预防性的，而不是弥补性的。这样就形成了一套理论，该理论定义了一个可以消除某一类错误（例如，接口错误）的系统。这套理论首先可以实现的技术就是根据功能分级来定义并构建可靠系统（Hamilton, 1986）。意识到通过系统定义的方式来解决主要问题（即可靠性）的好处之后，研究就会不断促进阿波罗哲学原理向前发展（以后也会继续）；这种哲学原理简单来说就是解决主要问题，并以同样的方式解决其他问题；也就是，利用语言机制内在特性来消除软件问题。这种原理最后带来的结果就是诞生了一种新的技术，即“事先预防”（Development Before The

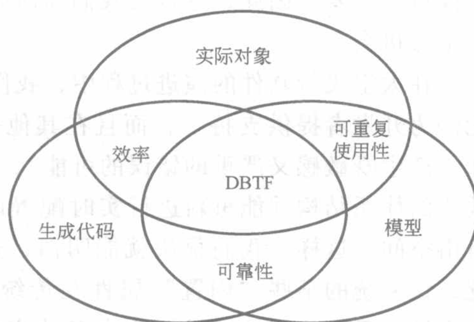


图 17-1 “事先预防”式开发技术

Fact，DBTF）式开发技术（见图 17-1）。在这种技术中，系统是根据预防性的特性来进行设计和构建的，这些特性综合了系统定义的所有方面，包括功能性网络和类型分级网络的内在综合（Hamilton and Hackler，in press；Hamilton，1994；Hamilton，1994；Keyes，2001，< http：//www. htius. com/ >）。

DBTF 是一种面向系统对象（System Oriented Object，SOO）的方法，这种方法基于控制的概念，这在其他软件工程模式中是不存在的。这种方法的理论基础是每一个系统都是一系列原理的集合（通常认为是真理），而且每一个 DBTF 系统的设计都是基于这些原理以及一系列常用的对象。每一种原理都定义了一种直接控制关系，而由这些原理定义的各种关系的组合就是控制。在其他事件中，这些原理就是指在一个对象的开发和操作阶段为对象建立了发起、输入输出、输入输出访问权限、错误检测及恢复以及调整各种控制关系。表 17-1 归纳了 DBTF 系统中对象的部分属性。

表 17-1　DBTF 系统中的面向系统对象特性

质量 *可靠性* *有效性* **可靠性** 可控制性 基于一组数学原理 域鉴定（有意、无意） 　分类（优先级和时间） 　访问权限：引入的对象（或者相关对象），引出的对象（或者相关对象） 　替换 格式 　　一致的、逻辑完整的 　　必要的、充分的 　　常用语义学基础 　　惟一的状态识别 无差错（基于“错误”的标准定义） 　　通常在正确的时间和正确的位置获得正确的答案 　　满足用户和开发者的各种意图 *处理意外事件* 可预见性 **有效性** *可重复使用性* 对操作和开发过程中的资源进行优化配置，使资源所占的时间和空间最小，并使资源与对象最匹配	**灵活性** 可改变，而不会产生边界效应 可演进性 持久性 *可靠性* 可扩展性 可分解及可重组能力 　　一个对象对多个对象：模块化、分解、实例 　　多个对象对一个对象：组合、实用的算子、综合、抽象化 可携带性 　　安全 　　多样性和不断变化的分层开发过程 　　开放式体系结构（实现、资源分配和单独执行） 　　插件程序或者不同模块的重新配置 不同组织、应用、功能性、人、产品的可改造性 **自动化操作** 可重复使用的最终形式 形式化，机械化，然后自动化 　　系统 　　系统的开发 　　实现系统自动开发的系统 **可掌握性、可集成性和可维护性** *可靠性* 时间不长 实际应用中的自然反应

（续）

可重复使用性

可掌握性、可综合性和可维护性

灵活性

自动化操作

常见定义

- 自然模块化
  - 自然分割（例如，根据资源体系结构来构建功能性体系结构）
  - 无声模型
  - 根据结构、性能和控制特性综合的对象
- 根据结构和性能的综合
- 机制类型
  - 功能图（将某个对象的功能与其他功能关联起来）
  - 对象类型图（将对象之间关联起来）
  - 功能和类型的构成

种类

- 相关性
  - 示例
  - 多态性
  - 父/子
  - 属性/行为
  - 现有特性/没有的特性
- 抽象概念
  - 封装
  - 替换
  - 相互关系，包含功能
  - 分类，包含各种级别
  - 形式，包含（对象类型和功能的）结构和性能
- 起源
  - 演绎
  - 推论
  - 遗传

处理意外事件

贯穿开发和操作过程来实现

不会对其他领域产生影响

检查错误并从意外事件中恢复

与异步、分布式、实时环境进行连接，并对其进行重新配置

- 持久特性、生成和消除
- 出现和消失
- 磨和容易性
- 参考
- 假设对象存在
- 实时和空间约束条件
- 描述方式
- *相关性、抽象性、起源*

提供用户友好定义

- 意识到一个用户的友好就是另一个用户的噩梦
- 隐藏不需要的细节（抽象化）
- 变量、用户选定的语法
- 自学
- 从一个常见语义基础中演变
- 常用定义机制

常用语义与其他所有实体之间的通信

尽可能简单定义，但不能过于简单

根据所有对象（和对象的所有属性）的综合来定义性能和结构以及贯穿其诞生、成长和消亡过程的变化（维护）

知识和实现的能力

- 定义
- 系统及其开发系统的开发
  - 分析
  - 设计
  - 实现
  - 实例化
  - 测试
  - 维护

注：表中的斜体字是指可重复使用的功能或性能。

进一步对 DBTF 方法进行研究后，我们就会明白，传统方法的根本问题在于它支持的是用户“在出现问题后进行修补”，而不是“在一开始就采用正确的方式来做事情”。在新的软件模式中，软件在开发之后首先不是进行软件测试来查

找错误，因为软件一开始就被定义成不允许存在错误，所谓的错误修改过程在软件的定义中就已经通过内置的语言属性来实现了。因此可以说，DBTF 方法创造的是一种通用的语义学，不仅可以用来定义软件系统，还可以用来定义常用的系统。

上述这些原理一旦理解之后，我们就会发现，可以将优秀设计的特性融合到语言中，来定义任何系统（不仅仅是软件系统），从而实现对优秀设计特性的重复使用。语言实际上是一种形式，它代表了系统的数学原理；语言（实际上是元语言）是 DBTF 的关键，其作用是帮助设计师降低设计的复杂度，同时提高设计师思想的清晰度，并最终将这些转换到最后的设计中。系统在定义时都是采用常用的系统语言，每个系统都可以合并到多元语言中去，并可以用来定义其他系统；采用这种合并语言定义的系统具有一些“理论上”的特性，这些特性控制着该系统的“命运”。DBTF 将系统的传统数学原理拓展了，并形成了独特的控制概念；基于这样的理论（DBTF），这种正规的但也很友好的语言已经融入到了时空物理学的自然描述中了。001AXES 逐渐成为了 DBTF 的正式通用系统语言，而 001 工具套件也成为了自动化工具。

在阿波罗工程中，重复使用定义方案可以节省软件开发的时间，也可以节省软件的空间，这种重复使用的概念成为了系统语言中高级语言语句类型的先驱。这种重复使用包括水平方向和垂直方向的重复使用，这就在 DBTF 环境中形成了灵活开放的体系结构。

理解测量机制的重要性及其对未来软件设计的影响是这个分析过程中的重点。这种设计方式在本质上代表了“可重复使用性”的初始状态，这种可重复使用性来自于阿波罗工程，其后就是各种改进的状态。可重复使用的对象包括那些有益于内部性能的功能，以及可以避免无益于内部性能事件出现的那些功能。那些从历史经验中得到的结论，今后还将在项目中接受实践的检验。有人曾经说过“当根据经验开发出来的东西变得与理论密切相关时，请不要感到惊讶”，DBTF 就是这样的一个例子。

17.7 “事先预防”理论

我们知道，数学方法是很难理解和应用的；而且在重要的系统和给定生命周期的系统中的应用也非常有限。在这方面，DBTF 之所以不同是因为其中的数学方法已经用来处理系统中的类（软件的一项属性）了。数学方法的形式体系“隐藏”在语言机制之下，而语言机制又来源于该形式体系，而且基于这种形式体系的技术已经得到了实用。

采用 DBFT 方法，所有的模型都可以根据其使用的语言来定义（而且采用自

动化环境来进行开发)，如 SOO。SOO 可以被开发环境中的所有其他对象以相同的方式共享，而且不会产生不明确的特性，其中开发环境包括：所有用户、模型和自动化工具。每个 SOO 都是由其他 SOO 构成的。SOO 的每个方面都是完整的，其中每个方面至少都是 SOO 中面向对象部分和面向功能部分以及面向时间部分的综合。因此，不能再说系统是面向对象的了，而应该说对象是面向系统的了；所有的系统都是对象，所有的对象也都是系统了。因此，很多以前采用传统方法不可能完成的事情，现在变得可能了；这一点看上去与传统方法完全相反，因为传统方法是以软件为中心的，而现在的 DBFT 是以系统为中心的。

DBFT 的自动操作为整个系统设计或软件开发生命周期内的系统设计师或软件开发者提供了遵循模式的各种方法，例如，测试方法。DBFT 模式开始预防的错误越多，测试的需求就越少，效率也就更高。"事先预防"测试方法是每一个开发步骤中固有的一部分，大多数错误都是通过这种固有属性来预防的。

与 UML 和 Java 一类只适用于软件的语言不同，001AXES 是一种系统语言，既适用于软件事件，也适用于任何给定的硬件事件。UML 是一种图形软件规范语言；而 001AXES 是一种图形系统规范语言，它们各适用于不同的目的。对于相同的事件，UML 用户必须采用某种编程语言进行编程，但是 001AXES 用户不用编程，而且系统完全可以根据 001AXES 所有的功能、对象和时间性能来进行定义和自动生成，甚至不用编写一串代码。UML 语言的目的是运行一个机器，而 001AXES 语言的目的是定义（如果有效，还要执行）一个系统软件或者其他东西。

由于 001AXES 是一种系统语言，它的语言语义学就比软件语言更加通用。例如，001AXES 定义的系统可以运行在计算机、人体、自动机器人或一个组织机构中；而软件语言中类的内容是基于计算机的，而计算机是运行软件的。

由于 001AXES 实质上是系统级的，它使用的都是系统中共有的内容，包括：软件、金融、政策、生物和物理系统。对于软件来说，001AXES 语言的语义学在软件中对应的是各种合适的结构，这些结构可能是一组底层的编程语言实现方式。不同于 DBFT 方法，这些更加传统的开发方法通常注重的是在系统及其软件的完整规范中采用分段式方法，这种方法稍稍改变了开发过程中的综合和自动操作过程。关于 001AXES 和 UML 更加详细的讨论，读者可以参考 Hamilton 和 Hackler 在 2000 年出版的著作。

17.8 设计及开发过程

每个 DBFT 都来自于各个传统方法步骤的合并，这些方法步骤用来解决传统系统工程和软件开发过程中的问题。每个 DBFT 系统都是根据内置质量、内置效

率和内置控制来定义的（类似于生物超有机体）。采用 DBFT 方法的系统，其设计过程不仅结合了数学的完整性，还结合了工程学的精确性，其主要目的就是促进“一开始就做正确的事情”的开发风格发展，同时避免“事后弥补”的传统方法。其自动操作在开发过程中必须考虑以下内容：在系统定义的早期阶段就防止错误的产生、在开发过程中实现对系统生命周期的控制以及实现高可靠系统中内在的重复使用功能。开发生命周期可以划分为一系列的阶段或者步骤，包括：通过正式的规范和分析进行需求和设计建模、基于相容的逻辑完整模型自动生成代码、测试执行以及仿真。

其中，第一步就是利用语言来定义模型。这个过程可以位于开发生命周期中的任何阶段，包括问题分析、操作环境和设计阶段。模型是自动进行分析的，以确保定义的正确性；这个分析过程包括对预防性的静态分析以及对用户意图特性的动态分析。下一步中，常见的源代码生成器就会在特定的语言和体系结构中，为选定目标环境中任何类型和任何大小的应用系统自动生成一个完全产品化和集成化的软件实现过程，该过程与模型是一致的。如果选定的目标环境已经配置好，那么生成器就会直接选用该目标环境；否则，生成器首先就要根据语言和体系结构进行配置。

由于 DBFT 方法是开放式体系结构，因此源代码生成器的配置可以停留在新的体系结构上（或者与外部环境的接口）；例如，配置成语言、通信数据包、互联网接口、数据库数据包或者特定操作系统；或者源代码生成器可以配置成与用户自己遗传代码之间的接口。一旦在一个新环境中配置好之后，现有系统就可以在新环境中自动重新生成。这种开放式的体系结构为用户改变需求或体系结构，或者从一种旧技术转换到新技术提供了更大的灵活性；同时这种方法也使自身成为了真正的基于构成的开发过程了。

系统生成之后就可以执行生成的系统了，如果生成的系统是一个软件，那么这个系统可以接受测试，以便进一步查找用户意图上的错误；测试完成之后，系统就可以进行操作了。应用系统会根据需求定义进行改变——而不是根据代码进行改变；目标体系结构也会根据生成器的环境配置进行改变（该环境根据模型产生一组实现方式）——而不是根据代码进行改变。如果实际系统是一个硬件或者人力系统，那么软件系统就可以作为实际系统的基础性仿真系统。一旦系统被开发出来，我们就可以对系统及其开发过程进行分析，以便掌握如何开展下一轮系统开发。

完整的综合或集成过程涉及了从系统到软件、从设计到需求、编码、测试和其他需求过程，以及各级和各层之间；这样，开发者就可以根据需求来追溯代码，或者反过来根据代码追溯需求。

假设自动操作具有以上功能，那么我们就不会对系统根据自身来定义并且根

据不断变化的体系结构和技术自动生成自己感到奇怪了。表17-2归纳了更加先进的预防性系统方法与传统方法之间的比较。

表17-2　预防性系统方法与传统方法之间的比较

传统方法(事后弥补)	DBFT(事先预防)
特别综合或者所有综合	综合
不一致的方法、对象、阶段、产品、结构、应用和环境	完整的生命周期:方法、对象、阶段、产品、结构、应用和环境
没有与软件综合的系统	与软件综合的系统
面向功能或面向对象	面向对象的系统:功能、时间、面向对象的综合
与应用没有综合的GUI	与应用综合的GUI
与软件代码没有综合的仿真	与软件代码综合的仿真
直到交货后才能保证性能	通过内置语言属性保证正确性
系统中无处不在的接口错误(占所有错误的75%)	没有接口错误
所有在系统实现之后才发现的错误	所有在系统实现之前就发现的错误
有些错误是人工发现的	所有的错误都是通过自动操作和静态分析发现的
有些错误是通过动态分析发现的	
有些错误无法发现	所有的错误都可以发现
模糊的需求、规范、设计…导致产生混乱、混淆和复杂性	明确的需求、规范、设计…消除混乱、混淆和复杂性
非正规或半正规语言	正规而且友好的语言
不同的阶段有不同语言和工具	所有阶段都具有共同的语言和工具
其他系统、软件采用不同的语言	所有的软件、硬件和其他系统都采用相同的语言
系统实现之后功能的完整性无法保证	系统实现之后功能的完整性可以保证
非灵活性:系统无法描述或无法演进	灵活性:系统可描述也可演进
受制于程序缺陷、需求产品、结构等等	开放式体系结构
从系统遗留问题中艰难地过渡	从系统遗留问题中平滑地过渡
在代码级水平上进行维护	在规范水平上进行维护
没有固有的重复使用性	固有的重复使用性
特定的对象才有重复使用性	每个对象都可作为重复使用的候选
专用化和重复使用性是相互排斥的	专用化会促进重复使用性的形成
自动化支持人工处理,但不能完成实际的工作	自动化完成实际的工作
大多数是人工完成的:文档整理、编程、测试生成、可描述性、综合	自动化编程、文档整理、测试生成、可描述性、综合
代码产生过程是有限的、不完整的、分段的、完全不相同的和低效率的	各种软件中,代码100%自动生成
产生未知变量,并根据其进行开发	根据001自身进行定义和生成
	所有赋值均为#1
系统浪费资金、错误较多	非常可靠的系统,在其开发过程中,效率非常高
高风险	低风险
没有成本效率	10:1、20:1、50:1…的成本节省/产出比
很难满足设计方案要求	在最短的时间内完成
很少有需要的,多半是不需要的	不多不少,正好符合需求

为了彻底掌握 DBFT 的概念，还需要一些相对专业的知识。DBFT 方法可以生成任何事件，这样在构建系统时就具有很强的可重复使用性。因此，我们很快就可以明白为什么再也不需要向语言中添加各种特征或者以特定的方式对开发应用过程进行改变，因为每一个新的属性都最终或本身就是来自于该方法的数学基础。

尽管 DBFT 方法成功应对了很多挑战，并解决了很多传统软件环境中的问题，但是 DBFT 方法在被主流用户接受之前经历了很长的时间和过程，因为它需要改变所有人的思维方式。这种现象在其他相关领域中也出现过。当计算机被发明时，所有的计算都是通过人工计算器进行的，因此，曾经有个说法就是连硬件先驱、数字设备公司的创始人 Ken Olson 都认为这个世界只需要 4 台计算机。因此，计算机的概念和广泛用途在经过了很长时间才开始流行。这种现象在更加先进的软件开发模式中也出现过。也就是说，软件在真正成为可以自动生成的产品（而不是像今天的某些传统开发环境中的手工生产）之前经历了很长时间。

17.9 选择正确的模式并实现自动操作

软件工程之所以会出现失败，是因为没有抓住要点，即不仅要选择正确的模式（很多模式都不是正确的），而且该模式必须是环境中的一部分，该环境可以提供内在的和完整的自动操作方法来解决手边的问题。也就是说，软件开发模式必须和一套完整的工具一起结合使用来生成模型，这就是设计和开发系统的方法。实质上就是，模式用来生成模型，而工具用来生成系统。

在很多情况下，通过技术上的投入来获得高效率系统的例子是很少的，很多实质性的问题都与组织机构构建其自动化系统的方式有关。尽管硬件性能得到了大幅提高，但是该组织机构却陷入了相同的方法论误区，该方法论在大型计算机开发面前就会产生风险。陈旧的方法论是不能简单地构建新系统的。

实际上，还有很多其他方面的变化。用户希望在其系统中能实现更多的功能，也希望系统能更加灵活；而且希望所有的问题都能通过这种新技术来解决，因此，这些系统必须是毫无错误的。

虽然生物超有机体具有内在的控制机制来确保其质量和效率；但到目前为止，硅超有机体还没有这种机制；因此，其效率是无法保证的。通常，解决关键问题的方法就是采用非传统的方法或者进行创新。如果使用新的方法，就必须创造出新的方法论或新的开发环境。成功的创新通常是从一个传统系统中的某个错误入手的。第一步就是真正认识到传统方法中的根本问题；然后根据如何预防它们来对其分类。下一步就是寻找这些问题的来源。重复上面的步骤并根据新的解决环境来寻找新的问题，这就是 DBFT 方法形成的原理。

采用 DBFT 方法时，系统设计与开发的各个方面都会与一种系统语言及其相

关的自动操作“综合”起来；在系统设计与开发的整个生命周期内，重复使用会一直贯穿始终。无论多复杂的对象都可以重复使用和综合。不同类型体系结构的环境配置也可以重复使用。一个新开发出来的系统也可以放心地重复使用，以进一步提高系统开发的效率。

这种预防性的方法可以支持用户应对当今软件开发环境中出现的各种挑战。但是，利用这种方法还有很多工作要做，即如何使这样一种技术更好地为我们服务。由于这种方法是基于不同的前提或假设（原理）的，因此，很多事情都会因为这一点而发生改变。以后我们还有更多的机会遇到新的研究项目和产品，其中包括了之前无法解决而现在可以得到解决的问题（得益于语言）。我们知道，软件开发从来都不会是相同的，因此，很多事情都不再需要了——事实上永远都不会存在了，例如生物系统自然选择过程中产生的那些现象。那些用来连接生命周期中各个阶段的技术已经变得陈旧了；而将编写源代码作为单独过程的技术也不再需要了，因为现在源代码都是根据需求规范来自动生成的。同样，认证技术也变得陈旧了。文档管理技术为我们提供了将需求及其变化直接融入需求规范数据库的方法。错误测试流程和查找工具也不再需要了，因为那些错误根本就不存在了。同样，支持人工编程的工具也不再需要了。

与采用传统方法进行开发相比，DBFT 开发出来的系统其效率更高（DOD，1992；Krut，1993；Ouyang，1995；Keyes，2000）。通过进一步分析，我们发现系统越大越复杂，其效率就越高——而相反的现象我们在传统的开发环境中曾经见过，这就因为其中大量重复使用了 DBFT 方法。系统越大，重复使用的机会越多；而重复使用得越多，系统的效率就不断提高。测量效率成为了一个相对的过程，也就是相对于最后开发出来的系统。

DBFT 开发环境中可重复使用性的利用在目前还是一个正处于不断发展的研究领域㊀（Hamilton，2003-2004）。例如，如何理解可重复使用类型与各种测量机制之间的关系，这就需要考虑对可重复使用性如何进行分类。一种方法就是根据重复使用性节省的时间来进行分类（节省的时间最后将演变成对成本和方案的影响）；以后将会出现更加合理有效的方法。我们知道重复使用性的类型越多，我们需要评估的整个系统成本的信息就越多。我们要记住，在软件开发的过程中，用来评估时间和成本的传统方法在采用预防性技术开发的系统中已经不再有效了。

㊀ 对常用指导性系统中常见软件界面的初步调查，该调查是根据 ARES 公司的合同进行的，该合同是 ARES 公司与 TACOM ARDEC 公司（Mr. Nigel Gray，项目经理）合同中的一部分。Mr. Gray 是 Common GUidance Common Sense（CGCS）系统、Deeply Integrated Guidance and Navigation “smart munitions”系统以及其他军用系统和航空系统应用的创始人。这些系统的常见软件界面可以在内部任务处理器编程时为用户提供一种常用的方法，而不用依赖于嵌入式操作系统。

这种高效率的产生还有其他的原因，例如采用预防性方法而节省下来的时间和资源以及某些不再需要的处理过程等等。因此，我们要学的和要做的东西就变少了，其中包括更少的分析、很少或几乎没有的实现过程、更少的测试、管理、存档、维护以及综合。这是因为这些领域中的大部分都是自动实现的，或者说这是正规系统语言的本质决定的。

设计师一旦意识到在设计和开发一个系统时不再需要那么多陈旧的工具了，那么系统模式就会发生变化。例如，如果采用了一种正式的语言来定义系统并综合系统所有的属性，那么各种各样的建模语言以及使用该语言的方法论（它们都只定义了系统某一方面的属性）都不再需要了，而且原本相互影响的各种技术也不再需要进行协调了。

软件作为一种比较新兴的技术领域，现在还处于不断发展的阶段。从一种传统软件开发环境到预防性软件开发环境的发展类似于从打字机到字处理器的发展，无论这种变化有多大，通常我们都需要学习新的解决问题的方法。但是，如同字处理器一样，每一点进步都会推动事物逐步向前发展。

总而言之，软件中方法论与技术的结合造就了软件成功的基础。软件在我们的社会中已经根深蒂固，它的成功或失败都将会深刻影响系统的运行以及政府和企业的成功。因此，今天关于系统工程和软件开发的各种决定都将会对以后产生长远的影响。

集体经验可以保证产品的质量和效率随着预防性系统特性的应用而逐步提高。与事实哲学背后的“迟做比不做好”哲理相比，预防性哲学可以尽早地解决（或者预防）给定的问题。这就是说，静态地发现问题比动态地发现问题好，通过系统的定义来预防问题发生更为可取，而且，还不用完全定义（或构建）系统。

重复使用一个可靠的系统总比重复使用一个不可靠的系统好，自动重复使用比人工重复使用好，系统本身的重复使用比自动的重复使用好。可以演进的重复使用性比不能演进的重复使用性好。最好是，重复使用那些接近真理的方法，然后再根据真理来指导这些方法的演进。这样的真理可以将常见的系统模式广泛应用到系统和软件中，并根据常用的语义学来统一它们的定义和理解。

DBFT 方法论中包含了预防性的系统方法论及其自动操作，其中系统方法可以通过语言来描述，而自动操作可以支持系统的各种应用；系统方法论和自动操作都是可以根据经验进行演进，从而进一步丰富这个理论，理论又进一步丰富语言，而语言又可以反过来促进自动操作向前发展。所有这些都可以同时用来设计系统和开发软件，这种原理就如同生物系统，其目标就是将来的系统可以继承当前系统中最好的东西。

名词解释

数据库管理系统（DBMS）：一种计算机程序，可以用来控制并提供对数据库的访问；DBMS 采用某种语言来控制 DBMS 提供的各种功能。例如，SQL 就是用来控制所有功能的一种语言，这些功能是指基于 DBMS 的相关体系结构而提供给用户的功能，包括：数据定义、数据恢复、数据处理、访问控制、数据共享以及数据完整性描述。

图形用户界面（GUI）：终端用户接口。通过该接口，开发出来的系统可以以可视化的方式与计算机高效地连接。图形用户界面提供了一系列直觉的、色彩丰富的图形机制，这种机制使得终端用户可以看见、更新并处理信息。

正规系统：根据一组著名的数学原理（或假设）定义的系统，因此该系统是基于数学原理的（例如，DBFT 系统是基于一组控制原理的）。该系统的部分属性之间是一致的，而且在逻辑上是完整的。如果一个系统中的原理或假设与其他原理或假设没有产生冲突，那么我们就认为该系统是一致的。如果系统方法中的假设完整定义了一组属性，那么我们就认为该系统在逻辑上完整的，而且这也确保了系统方法中的模型具有该组属性；而模型中的其他属性不一定能从该方法的假设中得到论证。一个逻辑上完整的系统具有一个语义学基础（例如，描述系统对象的方式）。根据 DBFT 系统的语义学，完整的系统意味着没有接口错误，而且系统是明确的，包含了必要的和充分的属性，并具有惟一的状态识别特性。

接口：对象、程序或系统之间的接合点，在该接合点处容易产生各种错误。软件可以通过接口与硬件、人力资源和其他软件进行连接。

方法论：软件构成中的一组流程、规则和结构。

测量机制（Metrics）：一系列的规则，这些规则可以用来测量质量和效率。

操作系统：用来管理并控制应用程序的程序。

软件体系结构：软件各组成部分之间的结构及其相互关系。

感谢

在此，要特别感谢 Jessica Keyes 的帮助性建议。

参考文献

[1] Boehm, B. W. 1981. *Software Engineering Economics.* Prentice-Hall, Englewood Cliffs, NJ.

[2] Booch, G., Rumbaugh, J., and Jacobson, I. 1999. *The Unified Modeling Language User Guide.* Addison-Wesley, Reading, MA.

[3] Department of Defense. 1992. *Software engineering tools experiment-Final report.* Vol. 1, Experiment Summary, Table 1, p. 9. Strategic Defense Initiative, Washington, DC.

[4] Goldberg, A. and Robson, D. 1983. *Smalltalk-80 the Language and its Implementation.* Addison-Wesley, Reading, MA.

[5] Gosling, J., Joy, B., and Steele, G. 1996. *The Java Language Specification*, Addison-Wesley, Reading, MA.

[6] Hamilton, M. 1994. Development before the fact in action. *Electronic Design*, June 13. ES.

[7] Hamilton, M. 1994. Inside development before the fact. *Electronic Design*, April 4. ES.

[8] Hamilton, M. (In press). *System Oriented Objects: Development Before the Fact.*

[9] Hamilton, M. and Hackler, W. R. 2000. Towards Cost Effective and Timely End-to-End Testing, HTI, prepared for Army Research Laboratory, Contract No. DAKF11-99-P-1236.

[10] HTI Technical Report, *Common Software Interface for Common Guidance Systems*, US. Army DAAE3002-D-1020 and DAAB07-98-D-H502/0180, Precision Guided Munitions, Under the direction of Cliff McLain, ARES, and Nigel Gray, COTR ARDEC, Picatinny Arsenal, NJ, 2003-2004.

[11] Harbsion, S. P. and Steele, G. L., Jr. 1997. *CA Reference Manual.* Prentice-Hall, Englewood Cliffs, NJ.

[12] Jones, T. C. 1977. *Program quality and programmer productivity.* IBM Tech. Report TR02. 764, Jan.: 80, Santa Teresa Labs., San Jose, CA.

[13] Keyes, J. The Ultimate Internet Developers Sourcebook, Chapter 42, *Developing Web Applications with* 001, AMACOM.

[14] Keyes, J. 2000. *Internet Management*, chapters 30-33, on 001-developed systems for the Internet, Auerbach.

[15] Krut, B. Jr. 1993. *Integrating* 001 *Tool Supply in the Feature-Oriented Domain Analysis Methodology* (CMN/SEI-93-TR-11, ESC-TR-93-188). Pittsburgh, PA: Software Engineering Institute, Carnegie-Mellon University.

[16] Lickly, D. J. 1974. *HAL/S Language Specification*, *Intermetrics*, prepared for NASA Lyndon Johnson Space Center.

[17] Lientz, B. P. and Swanson, E. B. 1980. Software Maintenance Management. Addison-Wesley, Reading, MA.

[18] Martin, J. and Finkelstein, C. B. 1981. *Information Engineering.* Savant Inst., Carnforth, Lancashire, UK.

[19] Meyer, B. 1992. *Eiffel the language.* Prentice-Hall, New York.

[20] NSQE Experiment: http://hometown. aol. com/ONeillDon/index. html, Director of Defense Research and Engineering (DOD Software Technology Strategy), 1992-2002.

[21] Ouyang, M., Golay, M. W. 1995. *An Integrated Formal Approach for Developing High Quality Software of Safety-Critical Systems*, Massachusetts Institute of Technology, Cambridge, MA, Report No. MIT-ANP-TR-035.

[22] Software Engineering Inst. 1991. *Capability Maturity Model.* Carnegie-Mellon University.

[23] Stefik, M. and Bobrow, D. J. 1985. Object-Oriented Programming: Themes and Variations. *AI Magazine*, Zerox Corporation, pp. 40-62.

[24] Stroustrup, B. 1994. *The Design and Evolution of C++*. AT&T Bell Laboratories, Murray Hill, NJ.

[25] Stroustrup, B. 1997. *The C++ Programming Language.* Addison-Wesley, Reading, MA.

[26] Yourdon E. and Constantine, L. 1978. Structured Design: Fundamentals of a Discipline of Computer Program and Systems Design, Yourdan Press, New York.

备注

读者还可以参考以下资料：

[1] Hamilton, M. and Hackler, R. 1990. 001: A rapid development approach for rapid prototyping based on a system that supports its own life cycle. *IEEE Proceedings*, First International Workshop on Rapid System Prototyping (Research Triangle Park, NC) pp. 46-62, June 4.

[2] Hamilton, M. 1986. Zero-defect software: The elusive goal. *IEEE Spectrum* 23 (3): 48-53, March.

[3] Hamilton, M. and Zeldin, S. 1976. Higher Order Software—A Methodology for Defining Software. *IEEE Transactions on Software Engineering*, vol. SE-2, no. 1.

[4] McCauley, B. 1993. Software Development Tools in the 1990s. AIS Security Technology for Space Operations Conference, Houston, Texas.

[5] Schindler, M. 1990. *Computer Aided Software Design*. John Wiley&Sons, New York.

[6] The 001 *Tool Suite Reference Manual*, Version 3, Jan. 1993. Hamilton Technologies Inc., Cambridge, MA.

第18章 神经网络与模糊系统

Bogdan M. Wilamowski

18.1 神经网络与模糊系统

随着电子元器件越来越新，性能越来越好，研究人员受到启发开始构建类似于人类神经系统的智能机器；这个目标最开始是由 McCulloch 和 Pitts（1943）在开发初等计算神经模型和 Hebb（1949）创建他的“学习规则”时提出来的。10年之后，Rosenblatt（1958）提出了“感知器”的概念。在20世纪60年代初，Widrow 和 Holf（1960，1962）开发出了智能系统，如 ADALINE 和 MADALINE。Nillson（1965）在他的著作《Learning Machines》中总结了那个时代的很多智能开发项目；而在 Mynsky 和 Paper（1969）的出版物中，关于人造神经网络方面的内容停止了一段时间，因为其中给出了一些让人泄气的结果；但是 Werbos（1974）却在不知不觉中构建了“后向传输”算法的数学基础。当前神经网络领域的快速发展开始于 Hopfield（1982，1984）的循环网络、Kohonen（1982）的无督管训练算法和 Rumelhart（1986）关于“后向传输”算法的描述等等。

18.2 神经单元

生物神经单元是一个复杂的结构，该结构负责接收数百个激励性和抑制性输入脉冲信号。这些输入脉冲在“隐蔽求和”过程中被赋以不同的加权系数（平均的）之后再进行求和。如果求和的值比门限值高，那么神经单元自身就会生成一个脉冲，该脉冲稍后被输入到相邻的神经单元中。由于输出脉冲是根据时间来求和的，因此，神经单元就会为更大的正值激励生成一个更高频的脉冲串。换句话说，加权输入端的求和值越高，神经单元产生脉冲的频率就越高。同时，每个神经单元在起始脉冲之后会沉寂一段时间，这段时间称为“不应期”。不应期可以更加准确地理解为一种现象；在该现象中，每当激励过后，门限值就会提高到一个很高的值，然后在一段时间内再逐渐下降。不应期为输出脉冲串频率设定了软上限。在生物神经单元中，信息是以调频脉冲串的形式进行传输的。

神经单元行为的描述涉及到了非常复杂的神经模型，该模型是理论上的模

型。McCulloch 和 Pitts（1943）指出，即使采用一个非常简单的神经模型也可以构建逻辑和存储电路。而且，这些具有门限值的简单神经单元通常比计算机中的典型逻辑门更加高效。McCulloch-Pitts 神经模型假设输入和输出信号只有 0 和 1 两个逻辑值，如果输入信号通过正的或者负的加权系数求和后得到了一个比门限值更大的值，那么神经模型的输出值就为 1；否则，神经模型的输出值就为 0。

$$T = \begin{cases} 1 & \text{如果} \quad net \geqslant T \\ 0 & \text{如果} \quad net < T \end{cases} \tag{18-1}$$

式中，T 是指门限值；net 是指所有输入信号的加权求和值；

$$net = \sum_{i=1}^{n} w_i x_i \tag{18-2}$$

McCulloch-Pitts 神经模型实现了 OR、AND、NOT 和 MEMORY 运算，如图 18-1 所示。注意到，OR 和 AND 门的结构是相同的。通过相同的结构，也可以实现其他逻辑功能，如图 18-2 所示。

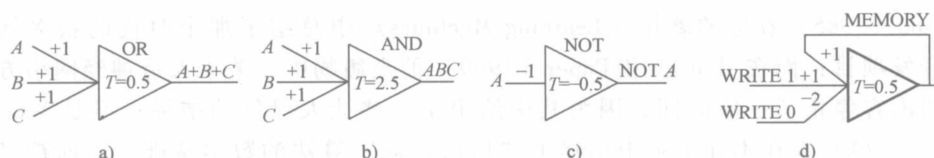


图 18-1 采用 McCulloch-Pitts 神经模型网络的 OR、AND、NOT 和 MEMORY 运算

a) OR b) AND c) NOT d) MEMORY

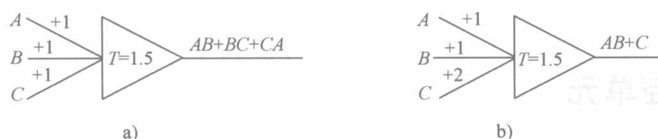


图 18-2 采用 McCulloch-Pitts 神经模型实现的其他逻辑功能

a) 多项乘积和 b) 单项乘积和

感知器模型具有类似的结构，其输入信号、加权值和门限值也可以是正的或负的。通常，在每个神经单元上都可以添加一个具有正或负加权值的附加输入常量，但不能使用可变的门限值，如图 18-3 所示。在这种情况下，门限值就通常设为 0，而加权求和值 net 就可

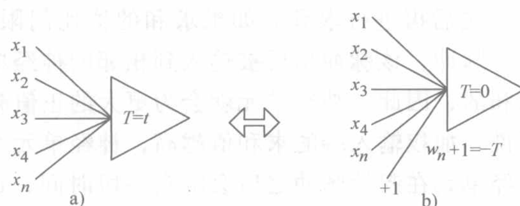


图 18-3 通过一个附加的加权值和一个 +1 常量输入实现的门限值

a) 门限值为 T 的神经单元 b) 修正后，门限值为 $T=0$，附加的加权值等于 $-T$ 的神经单元

以按照下面的公式来计算：

$$net = \sum_{i=1}^{n} w_i x_i + w_{n+1} \tag{18-3}$$

式（18-3）中，w_{n+1}的值与理想门限值相同，但符号相反。单层感知器已经被成功用来解决很多模式分类问题。单极性神经单元的固定门限触发函数由下式给出：

$$o = f(net) = \frac{\text{sgn}(net) + 1}{2} = \begin{cases} 1 & \text{如果} \quad net \geqslant 0 \\ 0 & \text{如果} \quad net < 0 \end{cases} \tag{18-4}$$

双极性神经单元的固定门限触发函数由下式给出：

$$o = f(net) = \text{sgn}(net) = \begin{cases} 1 & \text{如果} \quad net \geqslant 0 \\ -1 & \text{如果} \quad net < 0 \end{cases} \tag{18-5}$$

对于这些类型的神经单元，大多数的著名训练运算法则只能在单层网络中用来调整加权值。

多层神经网络通常采用持续触发函数，对于单极性神经单元：

$$o = f(net) = \frac{1}{1 + \exp(-\lambda net)} \tag{18-6}$$

对于双极性神经单元：

$$o = f(net) = \tanh(0.5\lambda net) = \frac{2}{1 + \exp(-\lambda net)} - 1 \tag{18-7}$$

这些连续触发函数适用于基于斜率的多层网络的训练，典型的连续触发函数如图 18-4 所示。当神经单元的输入端具有附加门限值时（见图 18-3b），参数 λ 可以从式（18-6）和式（18-7）中消去，神经单元响应的不合理性只能通过加权比例来控制。因此，就不必使用具有可变增益的神经单元了。

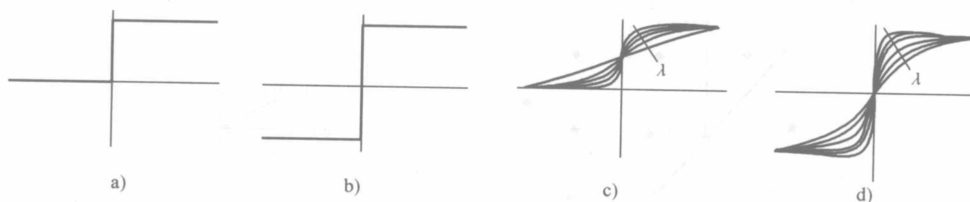


图 18-4　典型触发函数

a）固定单极性门限　b）固定双极性门限　c）连续单极性　d）连续双极性

注意到，即使是具有连续触发功能的神经模型也与实际的生物神经单元相差很远，实际的生物神经单元是根据调频脉冲串来工作的。

18.3 前馈神经网络

前馈神经网络中只允许一个方向的信号流，而且，大多数前馈神经网络是以层的形式构造的。图 18-5 给出了一个 3 层前馈神经网络的示例，该网络由输入节点、两个隐蔽层和一个输出层构成。

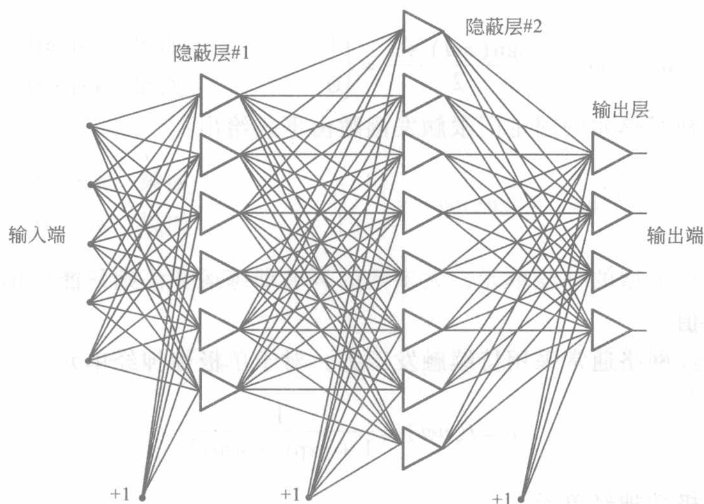


图 18-5 3 层前馈神经网络，有时也称为“后向传输网络”

一个简单的神经单元可以将输入模式分隔成两种类型，而且这种分隔是线性的。例如，在如图 18-6 所示的模式中，分隔线穿过了 x_1 和 x_2 轴，分别相交于 x_{10}和 x_{20}。这种分隔方法可以通过将神经单元中的加权值设为：$w_1 = 1/x_{10}$，$w_2 = 1/x_{20}$以及 $w_3 = -1$ 来实现。通常，对于 n 维网络，加权值为

$$w_i = \frac{1}{x_{i0}} \quad 其中，w_{n+1} = 1 \tag{18-8}$$

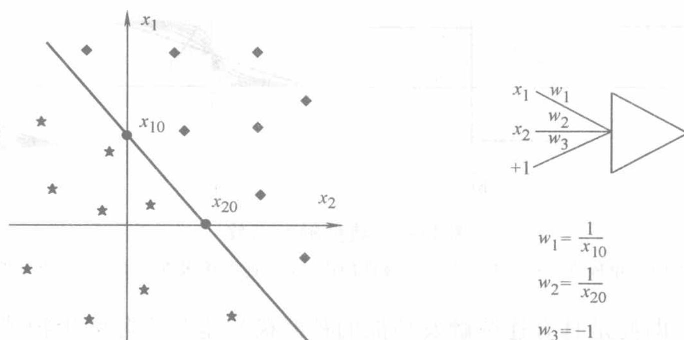


图 18-6 在 2 维空间中通过单个神经单元来线性分隔模式的特性

一个神经单元只能线性划分分隔模式。为了在 n 维输入空间中选择一个合适的区域，使用的神经单元必须多于 $n+1$ 个。如果有更多的输入端需要选择，那么输入（隐蔽）层中神经单元的数量必须对应地成倍增加。如果输入（隐蔽）层中神经单元的数量没有限制的话，那么所有的分类问题都可以通过 3 层网络来解决。图 18-7 给出了这样的神经网络的例子，该网络在一个 2 维空间对 3 个输入群进行分类。第 1 个隐蔽层中的神经单元负责在各输入群之间生成分隔线；第 2 个隐蔽层中的神经单元负责执行 AND 运算，如图 18-1b 所示；输出端神经单元负责为每一类执行 OR 运算，如图 18-1a 所示。神经单元的线性操作特性使得有些问题在神经网络中实现起来特别困难，例如专用 OR、多位的奇偶性计算、两个相邻螺旋上模式的分隔。

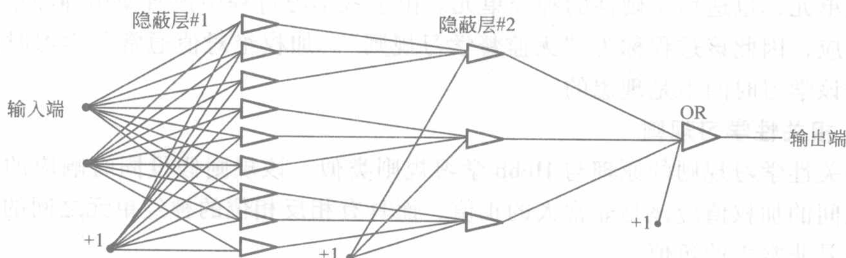


图 18-7 具有 2 个输入端的 3 层神经网络

注：这 2 个输入端用来将 3 个不同的神经输入群划分为一种类型，该网络可以广泛推广，而且可以用来解决所有的分类问题

前馈神经网络通常也可以用于多维输入变量与多维输出变量之间的非线性转换（映射）。理论上，如果神经网络在各隐蔽层中具有足够多的神经单元（输出层的大小由输出端所需神经单元的数量决定），那么任何输入-输出映射都是可能实现的。而实际上，这不是一件简单的任务。现在，还没有一个让人满意的方法来准确确定各隐蔽层中需要使用多少个神经单元，我们通常都是通过反复训练的方法来确定的。我们一般都知道，使用的神经单元越多，映射出来的形状就越复杂。换句话说就是，神经网络中的大量神经单元使其失去了推广使用的意义，而且这种神经网络中可能还要映射输入端的噪声。

18.4 神经网络的学习算法

类似于生物神经系统，人造神经系统中的加权值是在训练过程中进行调整的。人们开发出了各种学习算法，但是只有很少一部分适用于多层神经网络。其中，有些算法在神经单元中只使用逻辑信号，而其他则需要从输出端获取信息；有些算法需要一个监督员，该监督员知道给定模式中需要什么样的输出信号，而

其他无监督的算法则不需要这些信息。常用的学习规则描述如下。

1. Hebb 学习规则

Hebb（1949）学习规则是基于下面的假设：如果两个相邻的神经单元必须被同时触发或使能无效，那么与这些神经单元相关的加权值就应该增加；对于处于相反相位中的神经单元，其加权值就应该减小；如果信号之间没有相关性，那么加权值就应该保持不变。这种假设可以通过下面的公式来描述：

$$\Delta w_{ij} = cx_i o_j \tag{18-9}$$

式中，w_{ij}是指第 i 个神经单元与第 j 个神经单元之间的加权值；c 是指学习常量；x_i是指第 i 个输入端上的信号；o_j是指输出信号。

训练过程在开始时通常将所有的加权值设为 0，这种学习规则既适用于软性的神经单元，也适用于硬性的神经单元。由于该学习过程中没有使用神经单元的理想响应，因此该过程称为“无监督学习规则”。加权绝对值通常与学习时间呈正比，该学习时间不是理想的。

2. 相关性学习规则

相关性学习规则的原理与 Hebb 学习规则类似，该规则假设同时响应的神经单元之间的加权值应该是非常大的正值，而具有相反相位的神经单元之间的加权值应该是非常大的负值。

与 Hebb 学习规则不同的是，相关性规则是有监督的学习规则。理想的响应 d_j（而不是实际响应 o_j）用于加权值的变化计算，如下式所述：

$$\Delta w_{ij} = cx_i d_j \tag{18-10}$$

这个训练算法通常在开始时将加权值设为 0。

3. “中间形态”学习规则

如果输入矢量和加权值是标准化的，或者只包含二进制双极性数值（-1 或 1），那么当加权值和输入信号相同时，*net* 的值将会是一个很大的正值。因此，只有当加权值与信号不同时，加权值才需要改变，如下式所示：

$$\Delta w_i = c(x_i - w_i) \tag{18-11}$$

注意，加权值所需要的信息只能从输入信号中获得。因此，“中间形态”学习规则是一个非常局限的而且无监督的学习规则。

4. “优胜劣汰”（Winner Takes All，WTA）规则

WTA 是“中间形态”学习规则的一个修正版本，在该规则中，只有具有最高 *net* 值的神经单元的加权值才需要进行修正；而剩余的神经单元的加权值保持不变。有时，该规则的修正方式还包括下面这种，即对少部分同时具有最高 *net* 值的神经单元的加权值进行修正。尽管这是一种无监督规则，但是由于我们不知道什么是理想的输出，因此仍然需要一个“裁判”或“监督器”来找出一个具有最高 *net* 值的“优胜者”。WTA 规则（由 Kohonen 于 1982 年开发）通常用于

自动化集群和输入数据的统计特性提取。

5. “外部形态”学习规则

在“外部形态”学习规则中，要求与某个节点相关的各个加权值应该等于与这些加权值相关的神经单元的理想输出值，如下式所示：

$$\Delta w_{ij} = c(d_j - w_{ij}) \tag{18-12}$$

式中，d_j 是指理想神经单元输出值；c 是指很小的学习常量，该值在学习过程中会进一步减小。

这是一个有监督的训练过程，因为理想的输出值必须是已知的。“中间形态”学习规则和“外部形态”学习规则都是由 Grossberg（1969）开发的。

6. Widrow-Hoff LMS 学习规则

Widrow 和 Hoff（1960，1962）开发了一个有监督的训练算法，该算法允许对一个神经单元进行理想响应训练。该算法引出了 *net* 值与输出值之间的最小方差，如下式所示：

$$Error_j = \sum_{p=1}^{P} (net_{jp} - d_{jp})^2 \tag{18-13}$$

式中，$Error_j$ 是指第 j 个神经单元的误差；p 是指应用模式的数量；d_{jp} 是指当应用第 p 个模式时第 j 个神经单元的理想输出。

net 由式（18-2）给出。

该规则也称为“LMS”规则。通过计算式（18-13）的导数，并结合 w_{ij} 值，就可以得到加权值变化量的计算公式，如下所示：

$$\Delta w_{ij} = cx_i \sum_{p=1}^{P} (d_{jp} - net_{jp}) \tag{18-14}$$

注意，加权值变化量 Δw_{ij} 是每个单独应用模式变化量的和。因此，在每个单独模式应用之后就可以对加权值进行校正了，该过程称为“增量更新过程”；而“累积更新过程”则是在所有模式都应用之后对加权值进行校正时进行的。增量更新通常可以得到一个更快的解，但该解同时也对应用模式的阶数很敏感。如果选择的学习常量 c 很小，那么这两种更新方法可以得到相同的结果。LMS 规则适用于所有类型的触发函数，该规则强行使 *net* 值等于理想值。但有时，这也不是我们需要的。*net* 的值是多少并不重要，重要的是 *net* 的值是正还是负。例如，一个很大的 *net* 值，如果符号正确，将会产生正确的输出和很大的由式（18-13）中定义的误差，这可能就是推荐使用的解。

7. 线性衰退

LMS 学习规则要求成百上千或成千上万次地反复使用式（18-14），才能得到合适的解；而使用线性衰退规则，只需要一步就可以得到相同的结果。

假设一个神经单元中，输入模式的矢量符号为 $\boldsymbol{X}$，加权矢量为 $\boldsymbol{w}$，***net*** 值的

矢量就可以通过下面的公式来计算：

$$Xw = net \tag{18-15}$$

式中，X 是指 $(n+1)\times p$ 矩阵；n 为输入端的数量；p 为模式的数量。

注意到，输入模式的数量通常都会加 1，而且这个附加的加权值会对门限值产生影响（图 18-3b）。这种方法类似于 LMS 规则，它假设了一个线性触发函数，因此 net 的值应该等于理想输出值 d，如下所示：

$$Xw = d \tag{18-16}$$

通常，$p>(n+1)$，而且上面的方程式只有在最小均方差时才能解开。

根据矢量法则，该方程式的解为

$$w = (X^{T}X)^{-1}X^{T}d \tag{18-17}$$

如果采用传统的方法，p 个等式中有 $n+1$ 个是未知的，式（18-16）必须转换成 $n+1$ 个等式，其中有 $n+1$ 个是未知的，如下所示：

$$Yw = z \tag{18-18}$$

上式中，矩阵 Y 和矢量 z 由下式给出：

$$y_{ij} = \sum_{p=1}^{P} x_{ip}x_{jp} \quad z_i = \sum_{p=1}^{P} x_{ip}d_p \tag{18-19}$$

加权值由式（18-17）给出，或者通过对式（18-18）求解也可以得到。

8. 德尔塔（Delta）学习规则

LMS 方法假设线性触发函数的 $net=0$，求解等式得到的解有时与理想值相差甚远，如图 18-8 所示，图中是一个简单 2 维例子，其中 4 种模式分别属于两种类型。在通过 LMS 规则得到的解中，一种模式的分类是错误的。如果误差定义如下：

$$Error_j = \sum_{p=1}^{P} (o_{jp} - d_{jp})^2 \tag{18-20}$$

那么，由于 $o=f(net)$ 而且 net 由式（18-2）给定，因此误差相对于加权值 w_{ij} 的导数就是

$$\frac{\mathrm{d}Error_j}{\mathrm{d}w_{ij}} = 2\sum_{p=1}^{P} (o_{jp} - d_{jp}) \frac{\mathrm{d}f(net_{jp})}{\mathrm{d}net_{jp}} x_i \tag{18-21}$$

注意，该导数与触发函数的导数 $f'(net)$ 呈正比。因此，这类方法可能只适用于连续触发函数，而且该方法不能与固定触发函数（18-4）和（18-5）共同使用；在这方面，LMS 方法就更加通用一些。大多数连续触发函数的导数分别如下所示，对于单极性单元（式（18-6））为

$$f' = o(1-o) \tag{18-22}$$

对于双极性单元（式（18-7））为

$$f' = 0.5(1-o^2) \tag{18-23}$$

如果采用累积法，神经单元的加权值 w_{ij} 应该产生一定斜率的变化，如下

所示：

$$\Delta w_{ij} = cx_i \sum_{p=1}^{P} (d_{jp} - o_{jp}) f'_{jp} \tag{18-24}$$

在每个应用模式的增量训练中，

$$\Delta w_{ij} = cx_i f'_j (d_j - o_j) \tag{18-25}$$

加权值的变化应该与输入信号 x_i、理想与实际输出值的差 $d_{jp} - o_{jp}$ 以及触发函数 f'_{jp} 成正比。类似于 LMS 规则，加权值在增量和累积方法中也可以更新。与 LMS 规则不同的是，德尔塔（Delta）规则通常会得到一个非常接近于理想值的解。如图 18-8 所示，当使用德尔塔（Delta）规则时，所有 4 种模式的分类都是正确的。

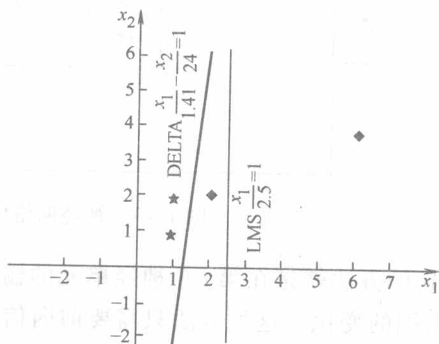


图 18-8　分别使用 LMS 和德尔塔（Delta）训练方法的结果比较

注：LMS 方法无法得到合适的解

9. 误差后向传输学习

德尔塔（Delta）学习规则适用于多层网络。通过采用类似于德尔塔（Delta）学习规则的方法，我们可以根据网络中的加权值来计算整体误差的斜率。如下所示：

$$\Delta w_{ij} = cx_i f'_j E_j \tag{18-26}$$

式中，c 是指学习常量；x_i 是指第 i 个神经单元输入端的信号；f'_j 是指触发函数的导数。

神经单元输出端的累积误差 E_j 如下所示：

$$E_j = \frac{1}{f'_j} \sum_{k=1}^{K} (o_k - d_k) A_{jk} \tag{18-27}$$

式中，K 是指网络输出端的数量；A_{jk} 是从第 j 个神经单元到第 k 个网络输出端的小信号增益，如图 18-9 所示。

后向传输误差的累积从输出层开始，累积误差可以一层一层计算直到输入层。从硬件实现的角度来看，这种方法是不切实际的。因此，取而代之的是，找出第 j 个神经单元的输入端到每个网络输出端的增益就容易得多，如图 18-9 所示。在这种方法中，加权值就是正确的，由式（18-28）给出：

$$\Delta w_{ij} = cx_i \sum_{k=1}^{K} (o_k - d_k) A_{jk} \tag{18-28}$$

注意，无论神经单元是否被排列在各层中，该公式都是通用的。一种获取增

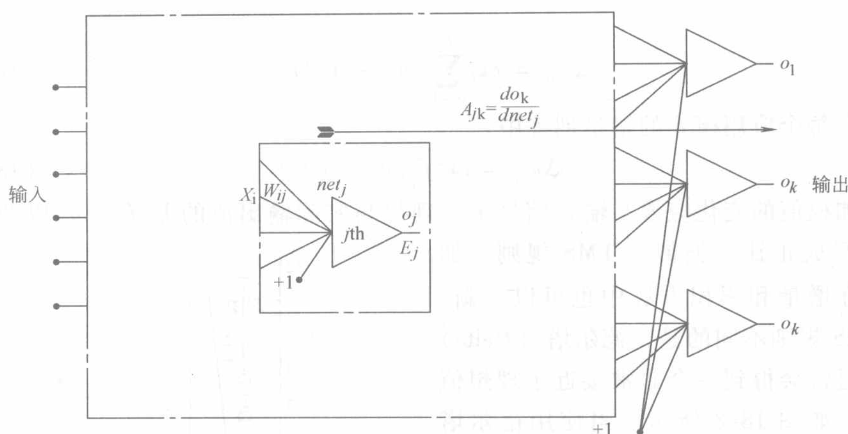


图 18-9 神经网络中增益计算的概念示例

益 A_{jk}的方法就是在第 j 个神经单元的输入端产生一个增量，然后观察第 k 个网络输出端的变化。这种方法只需要前向信号传输，而且在硬件实现中非常容易。另一种方法就是，计算通过每一层的增益，然后将各层的增益相乘得到总的增益。这种方法的计算量比误差后向传输算法中累积误差的计算量小或者差不多。

后向传输算法具有振荡倾向；为了使该算法过程尽量平滑，加权增量 Δw_{ij}可以根据 Rumelhart、Hinton 和 Wiliams（1986）公式来进行修正，如下所示：

$$w_{ij}(n+1) = w_{ij}(n) + \Delta w_{ij}(n) + \alpha \Delta w_{ij}(n-1) \tag{18-29}$$

或者根据 Sejnowski 和 Rosenberg（1987）公式来修正，如下所示：

$$w_{ij}(n+1) = w_{ij}(n) + (1-\alpha)\Delta w_{ij}(n) + \alpha \Delta w_{ij}(n-1) \tag{18-30}$$

式中，α 为动量。

当得到斜率的构成之后，后向传输算法的运算速度就可以大幅度提高，而且加权值可以沿着斜率的方向进行修正，直至达到一个最小值。该过程可以在每一步中进行，而且不存在大量的斜率计算。一旦在前面的斜率方向上获得了最小值，就要计算新的斜率构成。这个过程只适用于累积加权值调整。寻找沿斜率方向最小值的一种方法包含 3 个步骤，该过程在沿斜率的方向上寻找 3 个点的误差，然后使用抛物线近似法直接跳到最小值。快速学习规则采用了前面描述过的方法，最初是由 Fahlman（1988）提出来的，称为“QuickProp 算法”。

后向传输算法存在很多缺陷，这些缺陷会导致收敛过程变得非常慢。其中，最糟糕的是，在后向传输算法中，神经单元的学习过程几乎都会彻底萎缩，从而导致输出产生严重错误。例如，如果神经单元的输出值接近 +1，而理想输出应该是接近 -1，那么神经单元的增益 $f'(net) \approx 0$，而且错误信号无法后向传输，因此学习过程就变得没有意义了。为了克服这种困难，Wilamowski 和 Torvik

(1993) 提出了一种修正过的导数计算方法。在该方法中，导数是指一条直线的斜率，该直线连接了输出值的点和理想值的点，如图 18-10 所示。

$$f_{修正后} = \frac{o_{理想的} - o_{实际的}}{net_{理想的} - net_{实际的}} \tag{18-31}$$

注意，当误差较小时，式 (18-31) 的值就全部收敛于输出值点处的触发函数导数。如果系统的维数增加，那么局部区域出现多个最小值的几率就会下降。我们一般认为，前面描述过的现象是误差后向传输算法中产生收敛的主要原因，与局部区域最小值中的陷阱无关。

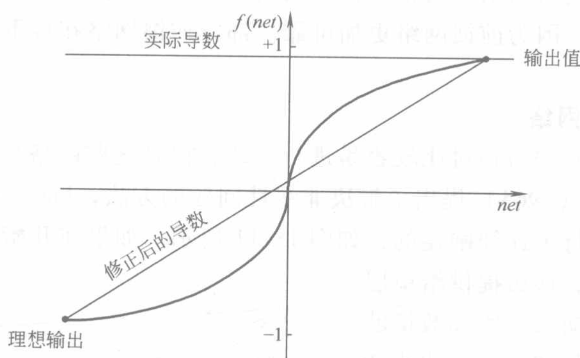


图 18-10　更快的误差后向传输收敛过程中的修正导数计算示例

18.5　特殊前馈网络

多层后向传输网络是一种常用的前馈网络，如图 18-5 所示。该网络由具有 S 形连续触发函数（见图 18-4c 和图 18-4d）的神经单元构成；大多数情况下，前馈网络只需要一个隐蔽层，该隐蔽层中的神经单元数量与待解决问题的复杂度成正比，具体数目通常是通过反复训练得到的。训练过程开始时，所有的加权值都被随机化为很小的值，然后用误差后向传输算法来求解。当学习过程无法收敛时，训练过程就会重复，同时采用一组新的随机选定的加权值。Nguyen 和 Widrow (1990) 为两层网络加权值的初始化提出了一种经验型的方法。在该方法的第 2 层中，加权值在 -0.5 ~ 0.5 的范围内随机选取；在第 1 层中，初始加权值的计算公式如下所示：

$$w_{ij} = \frac{\beta z_{ij}}{\|z_j\|};\ w_{(n+1)j} = \mathrm{random}(-\beta, +\beta) \tag{18-32}$$

式中，z_{ij}是 -0.5 ~ 0.5 范围内的随机数；缩放因子 β 由下式给出：

$$\beta = 0.7P^{\frac{1}{N}} \tag{18-33}$$

式（18-32）中，n 是指输入端的数量；式（18-33）中，N 是指第 1 层中隐蔽的神经单元的数量。这种类型的加权值初始化通常会使问题的求解过程变得更快。

为了得到后向传输网络中最合适的解，需要采用不同的网络结构和不同的初始随机加权值来进行反复训练。重要的是，训练后的网络必须获得一个概括性的特性，这就意味着训练后的网络应该能够正确处理那些没有进行训练的模式。因此，在训练过程中，训练模式中的部分数据通常会被移出，之后这些模式将用于验证。后向传输网络的结果通常存在很大的运气成分，这一点可以鼓励研究人员去开发前馈网络，因为前馈网络更加可靠。部分前馈网络在接下来的内容中将会进行描述。

1. 功能链接网络

单层神经单元网络相对比较容易进行训练，但是这些网络只能解决线性分隔问题。而 Nilsson（1965）提出了解决非线性问题的方法，Pao（1989）利用功能链接网络对其进行了详细阐述的，如图 18-11 所示。如果在开始确定功能时采用了非线性的概念，那么提供给单层神经单元网络的输入端实际数量就会增加。最简单的情况下，非线性单元是输入模式的更高阶形式。注意，功能链接网络也可以看做是一个单层网络，其中附加性输入数据通过非线性转换之后就不再是线性的了。单层网络的学习过程更简单，也更快。图 18-12 给出了一个采用功能链接网络来解决 XOR 问题的示例。注意，当使用功能链接方法时，这个复杂的问题就变成一个微不足道的问题了。功能链接网络的主要问题是选择合适的非线性单元不是一件容易的事情。但是，在很多实际例子中，我们很容易就可以预测出哪种输入数据的转换可以使问题线性化，这样就可以使用功能链接网络了。

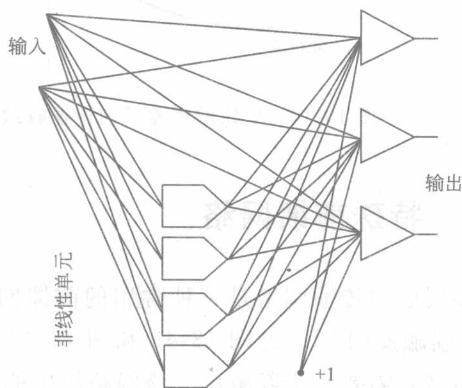


图 18-11　功能链接网络

2. 前馈型反向传输网络

“反向传输网络”最初是由 Hecht-Nilsen（1987）提出来的。在本节中，将讨论一个修订过的前馈型版本，该版本由 Zurada（1992）进行过的描述。如图 18-13 所示，该网络需要和输入模式同样多的，或者更准确一些，需要和输入群

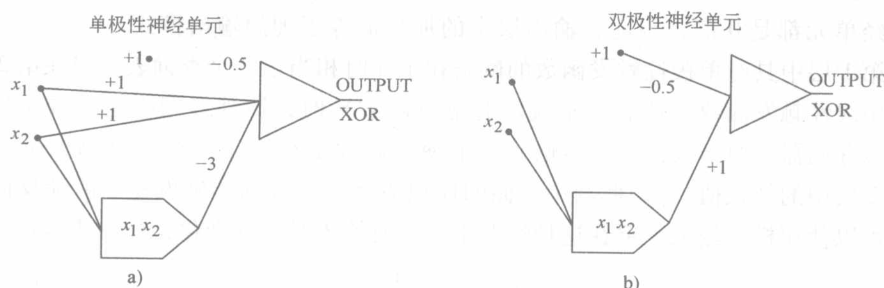


图 18-12　用功能链接网络来解决 XOR 问题

a）采用单极性信号　b）采用双极性信号

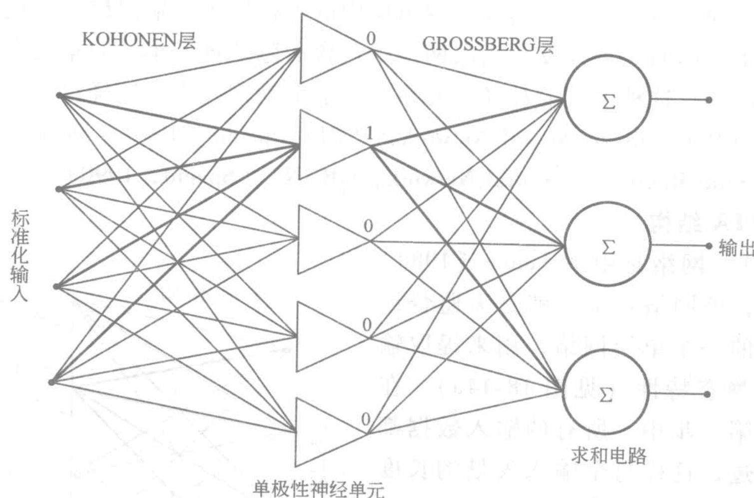


图 18-13　反向传输网络

同样多的隐蔽神经单元。第 1 层称为“Kohonen”层，该层中没有单极性神经单元，只有一个神经单元可以被触发，即优胜者。第 2 层称为“Grossberg 外部形态”层。Kohonen 层可以在无监督模式中进行训练，但这不是我们想要的。当使用二进制输入模式时，输入加权值必须严格等于输入模式，在该例子中，

$$net = \boldsymbol{x}^t \boldsymbol{w} = [n - 2HD(\boldsymbol{x}, \boldsymbol{w})] \tag{18-34}$$

式中，n 是指输入端的数量；$\boldsymbol{w}$ 是指加权值；$\boldsymbol{x}$ 是指输入矢量；$HD(\boldsymbol{x}, \boldsymbol{w})$ 是指输入模式和加权值之间的汉明距离（Hamming Distance）。

对于输入层中对存储模式起反应的神经单元，其门限值应该为

$$w_{(n+1)} = -(-n-1) \tag{18-35}$$

如果要求神经单元必须对存储的模式起反应，那么门限值应该被设为 $w_{(n+1)} = -[-n-(1+HD)]$，其中 HD 是指汉明距离，它定义了这种相似性的

范围。由于对于给定的输入模式，第1层中只有一个神经单元具有1值，而剩下的神经单元都是0值，因此，输出层中的加权值等于理想输出模式。

第1层中具有单极性触发函数的网络在工作时相当于一个查询表。当在第2层中使用线性触发函数（或者根本就没有触发函数）时，该神经网络也可看作是一个模拟存储器。由于模拟存储器输入端上使用的地址相当于一个二进制矢量（如同第2层中的加权值），因此其中存储的模拟值就可以得到准确恢复。前馈反向网络也可以使用模拟输入，但在这种情况下，所有输入必须标准化，如下所示：

$$\boldsymbol{W}_i = \hat{\boldsymbol{x}}_i = \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|} \tag{18-36}$$

反向网络很容易设计；隐蔽层中的神经单元数量等于模式（群）的数量；输入层中的加权值等于输入模式；输出层中的加权值等于输出模式。这种简单的网络可用于快速的原型开发。反向网络中包含的隐蔽神经单元数量通常比所需的数量多。但是，这种多余的神经单元在混杂的前馈网络中也同样存在，如随机性神经网络（Probabilistic Neural Network，PNN）Specht（1990）和广义回归神经网络（General Regression Neural Network，GRNN）（Specht，1990）。

3. WTA 结构

“WTA”网络是由 Kohonen（1988）提出来的，该网络基本上就是无监督训练法则中的一个单层网络，用来提取输入数据的静态特性（见图 18-14a）。在该网络的第一步中，所有的输入数据都被标准化过，这样每个输入矢量的长度都是相同的而且通常等于1（如式（18-36）所示）。神经单元的触发函数是单极性的，而且是连续的。学习过程开始时，加权值被初始化为一个随机值。在学习过程中，只有在输出端上具有最高值的神经单元（优胜者）上的加权值才可以改变，如下所示：

$$\Delta \boldsymbol{w}_w = c(\boldsymbol{x} - \boldsymbol{w}_w) \tag{18-37}$$

式中，$\boldsymbol{w}_w$ 是指优胜单元的加权值；$\boldsymbol{x}$ 是指输入矢量；c 是指学习常量。

通常，这种单层网络会排列成一个二维的形状，如图 18-14b 所示。图中的六边形可以保证神经单元之间产生很强

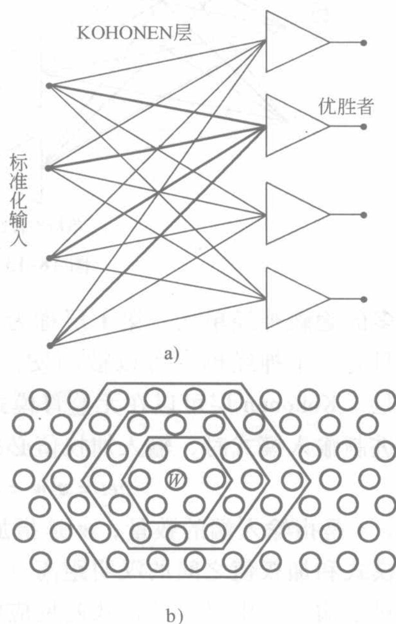


图 18-14 无监督训练模式中用于输入群筛选的 WTA 结构
a）网络连接 b）排列成六边形的单层网络

的相互作用。同样，算法也可以按照如下的方式进行修订，即不仅优胜单元中的加权值可以改变，而且相邻单元中的加权值也可以改变；同时，式（18-37）中的学习常量 c 也会随着与优胜单元之间距离的增大而减小。在这样一个无监督训练流程结束之后，Kohonen 层就可以组织数据进入输入群了。稍后，Kohonen 层的输出将会根据误差后向传输算法连接至单层或双层的前馈网络。这种 WTA 层中的初始数据组织方法通常会加快下一层中数据的传输。

4. 层叠相关性结构

层叠相关性结构是由 Fahlman 和 Lebiere（1990）提出来的，在开始构建该结构的网络时，需要添加一个单层神经网络和多个隐蔽神经单元，网络结构如图 18-15 所示。在每一个训练步骤中，都会添加一个新的隐蔽单元，而且该隐蔽单元的加权值会被调整，以使新的隐蔽单元的输出信号和网络输出端的剩余误差信号之间的相关性最大，该误差信号稍后将会被消除；相关性参数 S 必须最大化，如下所示：

$$S=\sum_{o=1}^{O}\left|\sum_{p=1}^{P}(V_{\mathrm{p}}-\overline{V})(E_{po}-\overline{E}_{o})\right| \tag{18-38}$$

式中，O 是指网络输出端的数量；P 是指训练模式的数量；V_p 是指新隐蔽神经单元的输出；E_{po} 是指网络输出的误差。

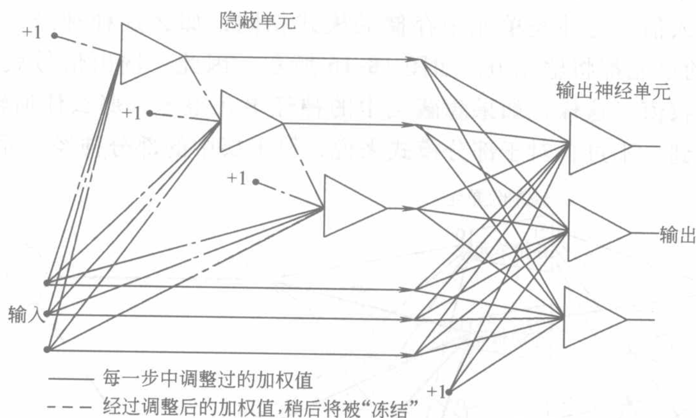


图 18-15 层叠相关性结构

$\overline{V}$ 和 $\overline{E}_o$ 分别是 V_p 和 E_{po} 的平均值。通过计算斜率 $\delta S/\delta w_i$，新的神经单元上的加权值调整如下：

$$\Delta w_i=\sum_{o=1}^{O}\sum_{p=1}^{P}\sigma_o(E_{po}-\overline{E}_o)f'_p x_{ip} \tag{18-39}$$

式中，σ_o 是新隐蔽单元的输出与网络输出之间相关性的符号；f'_p 是指模式 p 的触发函数的导数；x_{ip} 是指输入信号。

输出单元是采用“德尔塔（Delta）”或“QuickProp”算法进行训练的。每个隐蔽单元只训练一次，然后其加权值就会被“冻结”起来；当得到了满意的结果后，网络学习和构建过程就结束了。

5. 放射状功能网络

放射状网络的结构如图 18-16 所示，这种类型的网络通常只有一个隐蔽层，该隐蔽层中包含了一些特殊的神经单元；每个神经单元只输出接近存储模式的输出信号。第 i 个隐蔽单元的输出信号 h_i的计算公式如下：

$$h_i = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{s}_i\|^2}{2\sigma^2}\right) \tag{18-40}$$

式中，$\boldsymbol{x}$ 是指输入矢量；s_i是指存储模式，代表了第 i 个群的中心点；σ 是指群的半径。

注意，这种“神经”单元的性能与生物神经单元的性能有很大区别。在这种“神经”单元中，激励不是输入信号加权和的函数，而是计算输入端和存储模式之间的距离。如果该距离为零，那么神经单元将会输出一个最大的输出值，等于 1。这种神经单元可以组织部分模式并生成各种输出信号，这些信号是相似性的函数。这种神经单元的特征比后向传输网络中的神经单元的特征更有效。因此，由这种神经单元构成的网络效率就更高。

如果输入信号与神经单元中存储的模式相同，那么这种神经单元将会输出 1，而剩下的单元都将输出 0，如图 18-16 所示。因此，输出信号就等于触发神经单元的加权值。这样，如果隐蔽层中的神经单元很多，那么任何输入/输出映射都可以得到。不过，对于部分模式来说，第 1 层中的部分神经单元同时将会响

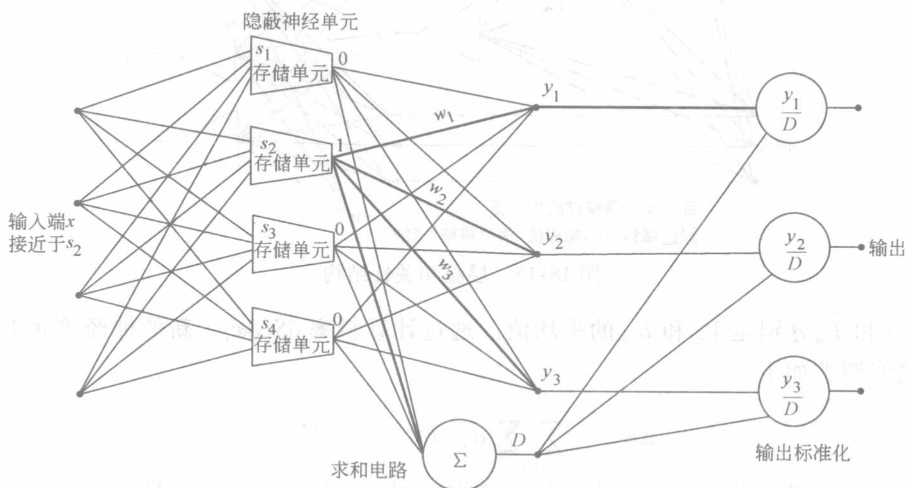


图 18-16 放射状功能网络典型结构

应一个非零的信号。如果近似值合适，那么隐蔽层中所有信号的总和应该等于 1。为了满足这一要求，输出信号通常必须标准化，如图 18-16 所示。

接下来就可以设计或训练放射状网络了。训练过程通常分两步进行；在第一步中，隐蔽层通常通过选择最佳的群描述模式来在无监督模式中进行训练。有一种类似于 WTA 结构中的方法可供使用。在这一步中，必须找到合适的群重叠半径 σ_i。

在训练过程的第二步中，误差后向传输算法只在输出层上执行。由于这是一个只对某一层有监督作用的算法，因此训练过程会非常快，可以比后向传输多层网络快 100 ~ 1000 倍。这一点使得放射状功能网络很有吸引力；而且，该网络很容易通过计算机进行建模；但是，其硬件的实现会很困难。

18.6　循环神经网络

与前馈神经网络不同，循环神经网络神经单元的输出可以连接到输入端，因此，网络中的信号可以连续循环。到目前为止，只有很少的循环神经网络被描述过。

1. Hopfield 网络

单层循环网络是由 Hopfield (1982) 进行分析的，如图 18-17 所示，该类型网络中包含单极性固定门限值神经单元，其输出值为 0 或 1。加权值由一个对称矩阵 W 给出，该矩阵主对角线上的元素都为 0 ($W_{ij}=0$, $i=j$)。系统的稳定性通常是通过"能量函数"来进行分析的，如下所示：

$$E=-\frac{1}{2}\sum_{i=0}^{N}\sum_{j=0}^{N}W_{ij}v_iv_j \tag{18-41}$$

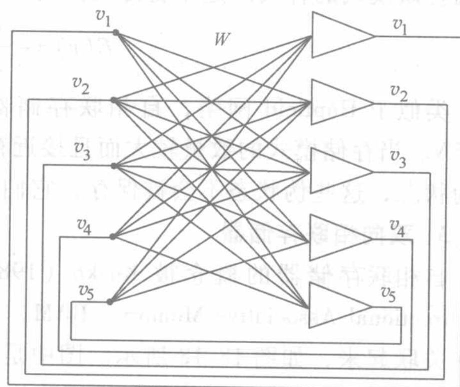


图 18-17　Hopfield 网络或自相联存储器

Hopfield 总结出了一定的规律，即在信号循环过程中，网络的能量 E 会逐渐下降，而系统的能量 E 会逐步收敛于一些稳定的点。尤其是当系统的输出值在异步模式中进行更新时，这个规律更加明显。这就意味着，对于给定的循环，只有一个随机的输出可以被修改成理想值。Hopfield 还总结出，通过修正后的 Hebb 规则来调整加权值，可以对那些系统能量收敛点进行规划，如下所示：

$$\Delta w_{ij} = \Delta w_{ji} = (2v_i - 1)(2v_j - 1)c \tag{18-42}$$

该类型网络中的存储器，其存储容量非常有限。根据经验，Hopfield 估计存储模式的最大值为 $0.15N$，其中，N 为神经单元的数量。

之后，能量函数的概念就被 Hopfield 扩展到单层循环神经网络中了，单层循环神经网络中的神经单元包含连续触发函数。这些类型的网络用来解决很多优化和线性规划问题。

2. 自相联存储器

Hopfield（1984）将他的网络概念扩展到了自相联存储器中。在相同的网络结构中，如图 18-17 所示，双极性固定门限值神经单元的输出值等于 -1 或 +1。在该类型的网络中，模式 $\boldsymbol{s}_m$ 根据自相关原则被保存在加权值矩阵 $\boldsymbol{W}$ 中，如下所示：

$$\boldsymbol{W} = \sum_{m=1}^{M} \boldsymbol{s}_m \boldsymbol{s}_m^{\mathrm{T}} - M\boldsymbol{I} \tag{18-43}$$

式中，M 是指存储模式的数量；I 是指单位矩阵。

注意，$\boldsymbol{W}$ 是一个对称方阵，其主对角线上的元素都等于 0（$W_{ij}=0$，$i=j$）。根据修订后的式（18-42），新的模式可以被添加到存储器中，也可以从存储器中去除掉。当这一类存储器通过强制遵循网络的初始状态而被应用到二进制双极性模式中时，在信号循环结束之后，网络将会收敛于最近（最相似）的存储模式或者该模式的补码。这个稳定的点将处于最接近最小能量的地方，如下所示：

$$E(v) = -\frac{1}{2}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{v} \tag{18-44}$$

类似于 Hopfield 网络，自相联存储器的存储容量非常有限，约为 $M_{\max} = 0.15N$。当存储模式的数量较大而且接近存储容量时，网络就会倾向于收敛于各种伪状态，这些伪状态不会被保存，它们是能量函数中额外的一些最小值点。

3. 双向相联存储器

自相联存储器的概念被 Kosko（1987，1988）扩展到了双向相联存储器（Bidirectional Associative Memory，BAM）中，双向相联存储器可以将模式 $\boldsymbol{a}$ 和模式 $\boldsymbol{b}$ 关联起来，如图 18-18 所示。图中是一个双层网络，其中第 2 层的输出与第 1 层的输入相连。第 2 层的加权值矩阵为 $\boldsymbol{W}^{\mathrm{T}}$，第 1 层的加权值矩阵为 $\boldsymbol{W}$。矩形加权值矩阵 $\boldsymbol{W}$ 是交叉相关矩阵的和，如下所示：

$$\boldsymbol{W} = \sum_{m=1}^{M} \boldsymbol{a}_m \boldsymbol{b}_m \tag{18-45}$$

式中，M 是指存储模式的对数；$\boldsymbol{a}_m$ 和 $\boldsymbol{b}_m$ 为存储的矢量对。

如果节点 a 和 b 在初始化时其中一个矢量类似于存储的矢量，那么在信号循环结束之后，存储的模式 $\boldsymbol{a}_m$ 和 $\boldsymbol{b}_m$ 都将会被恢复。与自相联存储器类似，BAM 的

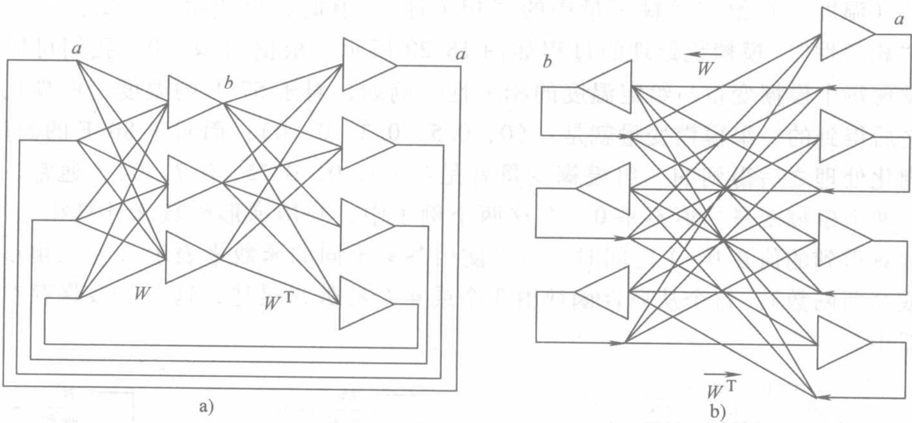


图 18-18　双向自相联存储器

a）形状像一个双层网络的信号循环电路　b）形状像一个双层网络的双向信号流

存储容量也非常有限，而且存在一定的存储器退化问题。BAM 概念也可以扩展到 3 个或更多的矢量相关性中。

18.7　模糊系统

神经网络的主要应用与 n 维输入变量和 m 维输出变量之间的非线性映射相关。控制系统中通常需要这样的功能；在控制系统中为了得到特定的测定变量，必须生成某些控制变量。另外一种将一组变量转换成另一组变量的非线性映射方法称为“模糊控制器”。模糊控制器的运算原理与神经网络有很大区别，模糊控制器的功能框图如图 18-19 所示。在模糊控制器的第一步中，模拟输入被转换成一组模糊变量；在这一步中，系统为每个模拟输入信号生成了 3 ~ 9 个模糊变量。每个模糊变量具有一个 0 ~ 1 之间的模拟值。在下一步中，使用模糊逻辑电路来输入模糊变量，并生成一组输出变量。在最后一步中（称为“去模糊处理”），从一组输出模糊变量中生成一组输出模拟变量，这一组模拟变量将作为控制变量。



图 18-19　模糊控制器的功能框图

1. 模糊化处理

模糊化处理的目的是将模拟输入变量转换成一组模糊变量。为了得到更高的精确度，将使用很多模糊变量。为了详细阐述模糊化处理过程，假设输入变量为温度，而且温度变量被编码成 5 个模糊变量：冷、凉、正常、暖和、热。每一个模糊变量都获得一个 0 ~ 1 之间的数值，称为模拟输入

变量（温度）在给定模糊变量中的“相关性”。有时，也使用“从属性”来代替“相关性”。模糊化处理的过程如图 18-20 所示。根据图 18-20，我们可以准确发现每个模糊变量与给定温度的相关性。例如，对于 57°F 的温度，模糊化处理之后得到的一组模糊变量就是：（0，0.5，0.2，0，0）；而对于 80°F 的温度，模糊化处理之后得到的一组模糊变量就是：（0，0，0.25，0.7，0）。通常，只有一两个模糊变量的值不是 0。在这两个例子中，利用梯形函数来计算相关性。当计算得到的值在 0～1 之间时，可以使用各种不同的函数来表示（如三角形函数或高斯函数）。每个从属性函数由 3 个或 4 个参数来描述，这些参数保存在存储器中。

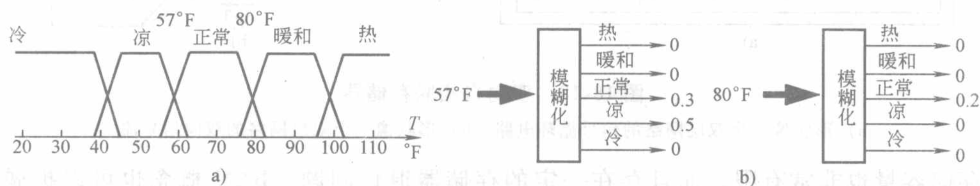


图 18-20　模糊化处理过程

a）模糊化和去模糊化过程的典型相关性函数　b）将温度变量转换成模糊变量的示例

为了正确设计模糊化的各个步骤，我们必须遵循以下的规则：

1）输入模拟变量的每个点都应该属于至少一个、至多两个从属性函数。

2）对于层叠函数，两个从属性函数的和不能大于 1。这就意味着层叠过程不能穿过最大值点（1）。

3）为了得到更高的精确度，应该使用更多的从属性函数。但是，过于密集的函数会导致频繁的系统反应以至系统出现不稳定。

2. 按规则赋值

与布尔逻辑相反，在模糊逻辑中，所有的变量都可以是 0～1 之间的任何值；而布尔逻辑中，变量只能是二进制值。模糊逻辑由与布尔逻辑相同的基础 ∧ - AND、∨ - OR 和 NOT 算子构成，如下所示：

$$A \wedge B \wedge C \Rightarrow \min\{A,B,C\}$$ ——A,B,C 中的最小值；

$$A \vee B \vee C \Rightarrow \max\{A,B,C\}$$ ——A,B,C 中的最大值；

$$A \Rightarrow 1\ 1-A$$ ——1 减去 A 的值

例如，$0.1 \wedge 0.7 \wedge 0.3 = 0.1$；$0.1 \vee 0.7 \vee 0.3 = 0.7$；$\overline{0.3} = 0.7$。这些规则称为“Zadeh” AND、OR 和 NOT 操作算子（Zadeh，1965）。注意，这些规则也适用于经典的二进制逻辑。

模糊映射表给出了模糊规则的具体描述，与给定系统中的情况一致。假设一个简单的系统具有两个模拟输入变量 x 和 y，以及一个输出变量 z。模糊系统设

计的目标就是将变量 z 转换成 $f(x, y)$ 函数。模糊化处理之后，模拟变量 x 由 5 个模糊变量表示：x_1、x_2、x_3、x_4、x_5，而模拟变量 y 由 3 个模糊变量 y_1、y_2、y_3来表示。假设模拟输出变量由 4 个模糊变量表示：z_1、z_2、z_3、z_4；那么设计过程中的关键就是为输入模糊变量的所有组合设置合适的输出模糊变量 z_k，如图 18-21 中的映射表所示。设计师必须指定很多规则，例如如果输入变量由模糊变量 x_i 和 y_j表示，那么输出变量就应该由模糊变量 z_k 表示。一旦模糊变量映射表指定以后，模糊逻辑计算就可以分为两个步骤了。首先，模糊映射表中的每个格子中都填入中间模糊变量 t_{ij}，该变量由 AND 运算得到，$t_{ij} = \min\{x_i, y_j\}$，如图 18-21b 所示。这一步与给定系统中的理想规则无关。在第二步中，OR（max）算子用来计算每个输出模糊变量 z_k。在图 18-21 给出的例子中，$z_1 = \max\{t_{11}, t_{21}, t_{31}, t_{41}, t_{51}\}$，$z_2 = \max\{t_{13}, t_{31}, t_{42}, t_{52}\}$，$z_3 = \max\{t_{22}, t_{23}, t_{43}\}$，$z_4 = \max\{t_{32}, t_{34}, t_{53}\}$。注意，这些变量的形式取决于图 18-21a 中模糊映射表给出的规范。

	y_1	y_2	y_3
x_1	z_1	z_1	z_2
x_2	z_1	z_3	z_3
x_3	z_2	z_4	z_4
x_4	z_1	z_2	z_3
x_5	z_1	z_2	z_4

a)

	y_1	y_2	y_3
x_1	t_{11}	t_{12}	t_{13}
x_2	t_{21}	t_{22}	t_{23}
x_3	t_{31}	t_{32}	t_{33}
x_4	t_{41}	t_{42}	t_{43}
x_5	t_{51}	t_{52}	t_{53}

b)

图 18-21 模糊映射表
a）模糊规则表 b）使用中间变量 t_{ij}的映射表

3. 去模糊处理

按照模糊规则赋值的结果是，每个模拟变量都由多个模糊变量来表示；而去模糊的目的就是获得模拟输出值。去模糊处理可以通过从属性函数实现，类似于图 18-20 中的模糊处理过程。在第一步中，根据规则赋值得到的模糊变量用来修正从属性函数，如下式所示：

$$\mu_k^* = \min\{\mu_k(z), z_k\} \tag{18-46}$$

例如，如果输出模糊变量为：0，0.2，0.7，0，0，那么修正后的从属性函数的形状如图 18-22 中的细线所示；变量 z 的模拟值是修正后从属性函数 μ_k^* 的“重心”，如下所示：

$$z_{模拟} = \frac{\left(\sum_{k=1}^{n}\int_{-\infty}^{+\infty}\mu_k^*(z)z\mathrm{d}z\right)}{\left(\sum_{k=1}^{n}\int_{-\infty}^{+\infty}\mu_k^*(z)\mathrm{d}z\right)} \tag{18-47}$$

如果输出从属性函数 $\mu_k(z)$ 的形状都相同，那么上面的公式就可以简化为

$$z_{模拟} = \frac{\left(\sum_{k=1}^{n} z_k zc_k\right)}{\left(\sum_{k=1}^{n} z_k\right)} \tag{18-48}$$

式中，n 是指 $z_{模拟}$ 输出变量的从属性函数的数量；z_k 是指从规则赋值得到的模糊输出变量；zc_k 是指与第 k 个从属性函数的中心对应的模拟值。

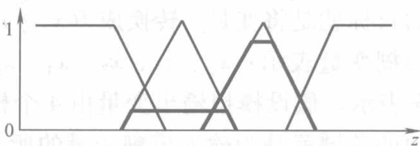


图 18-22　去模糊处理过程

式（18-47）对于基于简单微控制器的系统来说过于复杂了，因此，从实际出发，式（18-48）使用得更多。

18.8　设计实例

假设要设计一个简单的洒水装置模糊控制器。洒水时间是湿度和温度的一个函数。温度有 4 个从属性函数，湿度有 3 个，而洒水时间则有 3 个从属性函数，如图 18-23 所示。根据经验，可以得到模糊映射表，如图 18-24a 所示。

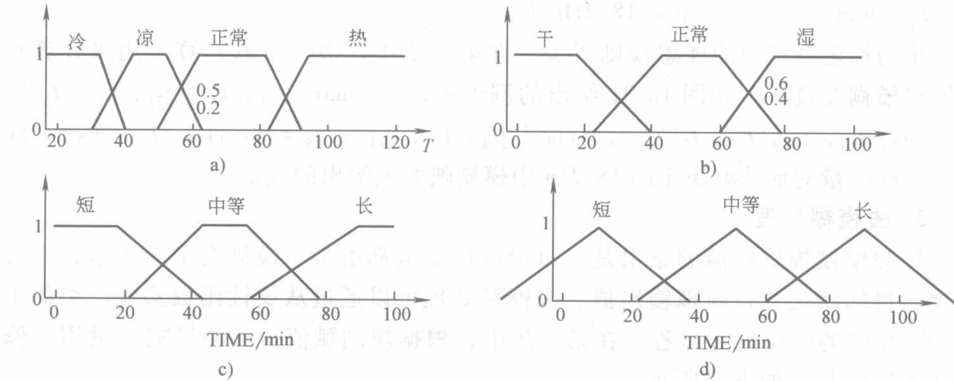


图 18-23　从属性函数示例

a）和 b）为输入变量的从属性函数　c）和 d）为输出变量的两种可能的从属性函数

假设温度为 60°F，湿度为 70%。利用温度和湿度从属性函数，可以得到如下的模糊变量：温度模糊变量为（0，0.2，0.5，0），湿度模糊变量为（0，0.4，0.6）。根据取小操作，模糊映射表中可以填入温度模糊变量，如图 18-24b

	干	正常	湿
冷	M	S	S
凉	M	M	S
暖和	L	M	S
热	L	M	S

a)

		干	正常	湿
		0	0.4	0.6
冷	0	0	0	0
凉	0.2	0	0.2	0.2
暖和	0.5	0	0.4	0.5
热	0	0	0	0

b)

图 18-24　模糊映射表设计示例

a）模糊规则　b）模糊温度变量

所示。注意，在表中只有4个小格为非零值。根据模糊规则（见图 18-24a），取大操作可以用来获得模糊输出变量：短→$O_1 = \max\{0, 0, 0.2, 0.5, 0\} = 0.5$；中等→$O_2 = \max\{0, 0, 0.2, 0.4, 0\} = 0.4$；长→$O_3 = \max\{0, 0\} = 0$。如果利用式（18-47），并根据图 18-23c，最后可以得到洒水时间为 28min；如果使用简化的方法，并根据式（18-46）和图 18-23d，最后可以得到洒水时间为 27min。

18.9　遗传规则方法

人造神经网络的成功激励了研究人员去寻找其他自然模式。进化过程中的遗传力量可以用来实现这种复杂的系统，例如人体系统。自然进化过程中的遗传规则可以更好地解决复杂的问题。遗传规则的基础是由 Holland（1975）和 Goldberg（1989）发现的。在人出生之后，每个遗传过程中将反复进行选择、交叉繁殖、基因突变等各个步骤。在这些过程中，某些符号串（称为“染色体”）就会朝着更好的方向发展。遗传规则方法在开始时是进行编码和初始化操作。遗传规则中所有的重要步骤都可以通过一个简单的例子来解释，即寻找函数 $(\sin^2 x - 0.5x)^2$ 的最大值，其中 x 的范围为 0 ~ 1.6，如图 18-25 所示。注意，在该范围内，函数在 $x = 1.309$ 处具有一个全局的最大值，在 $x = 0.262$ 处有一个局部最大值。

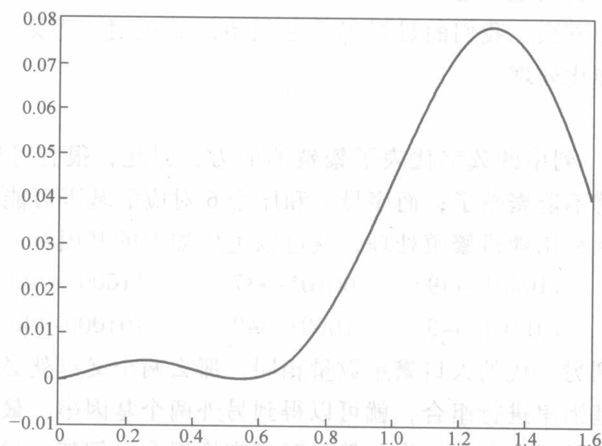


图 18-25　函数 $(\sin^2 x - 0.5x)^2$ 曲线图

1. 编码和初始化

首先，变量 x 必须表示成一串符号。符号串越长，过程通常收敛得越快，因此，一个符号串中使用的符号数越少越好。尽管这种符号串中包含的可能是任意符号，但是通常使用的是二进制符号 0 和 1。在例子中，我们使用 6 位的二进制

码来编码，因此得到了一个十进制的值40x。编码开始时生成了一组随机的原始数据，如表18-1所示。

表18-1 原始数据

符号串序号	符号串	十进制值	变量值	函数值	在总体所占的份数
1	101101	45	1.125	0.0633	0.2465
2	101000	40	1.000	0.0433	0.1686
3	010100	20	0.500	0.0004	0.0016
4	100101	37	0.925	0.0307	0.1197
5	001010	10	0.250	0.0041	0.0158
6	110001	49	1.225	0.0743	0.2895
7	100111	39	0.975	0.0390	0.1521
8	000100	40	0.100	0.0016	0.0062
总计				0.2568	1.0000

2. 选择和繁殖

在所有基因中选择最佳的基因是遗传规则中一个非常重要的步骤。有很多种不同的方法可以用来对每个个体基因进行排列等级。在这个例子中，排列函数已经给出来了。其中，最高等级中包含了序号为6的基因，而最低等级中包含了序号为3的基因。等级越高的基因，可以繁殖的几率越大。每个基因繁殖的能力可以从其在所有目标函数值的总和中所占的比例可以看出来，如表18-1所示。注意，为了使用这种方法，我们的目标函数必须始终是正值。否则，如果是负值，首先必须进行标准化处理。

3. 繁殖

表18-1最后一列中的数字代表了繁殖的能力，因此，很有可能序号3和序号8对应的基因就不能繁殖了；而序号1和序号6对应的基因可能具有两个或更多的复制品。如果采用随机繁殖处理，就可以生成如下的基因对：

101101→45　110001→49　100101→37　110001→49

100111→39　101101→45　110001→49　101000→40

如果从一代到另一代的人口繁殖数量相同，那么两个父母就必须繁育两个孩子。通过将两个基因串进行组合，就可以得到另外两个基因串。最简单方法就是将父母中每个人的基因串各分一半，然后互相交换组合。例如，父母的基因串为010100和100111，那么他们子女的基因串就可以是0101111和100100，这个过程称为“交叉繁殖”。最后得到的基因对为

101111→47　110101→53　100001→33　110000→48

100101→37　101001→41　110101→53　101001→41

一般来说，基因串不必各分一半来进行组合。通常，只要将父母之间选定的

基因位相互交换就足够了，惟一重要的是基因位置不能发生变化。

4. 基因突变

在进化的过程中，繁殖过程是通过基因突变来提高质量的。除了父母天生的特性之外，子女会获得一些新的随机特性，这个过程就称为“基因突变”。在大多数情况下，基因突变会诞生出低等级的孩子，这在繁殖过程中是需要消除的。但是，有时基因突变也可能产生一个具有更好特性的个体，这就防止了繁殖过程的退化。在遗传规则中，基因突变通常扮演一个次要的角色。对于高水平的基因突变，其过程类似于随机模式生成过程，而且这样一个搜寻规则的效率是很低的。基因突变率通常在正常情况下低于1%。在该例子中，基因突变的概率等于给定模式中随机位变化的概率。最简单的例子是，如果采用的是短基因串和很少的基因数量，那么典型的基因突变概率为0.1%，而在基因突变过程中，模式实际上是不会改变的。该例子的第二代人如表18-2所示。

表18-2 第2代人口

符号串序号	符号串	十进制数值	变量值	函数值	在总体所占的份数
1	010111	47	1.175	0.0696	0.1587
2	100100	37	0.925	0.0307	0.0701
3	110101	53	1.325	0.0774	0.1766
4	010001	41	1.025	0.0475	0.1084
5	100001	33	0.825	0.0161	0.0368
6	110101	53	1.325	0.0774	0.1766
7	110000	48	1.200	0.0722	0.1646
8	101001	41	1.025	0.0475	0.1084
总计				0.4387	1.0000

注意，第二代中的两个相同高等级成员非常紧接近于 $x = 1.309$ 处的解决方案。第三代随机选择的父母基因串为

010111→23　110101→53　110000→48　101001→41
110101→53　110000→48　101001→41　110101→53

他们生出来的孩子的基因串为

010101→21　110000→48　110001→49　101101→45
110111→55　110101→53　101000→40　110001→49

在第三代人中，最好的结果与第二代相同。通过仔细观察第二代到第三代的所有基因串，我们可以发现，通过交叉繁殖（基因串通常是对半分开），无论诞生多少代人，最理想的解决方案110100→52都很难达到，这是因为第二代中没有人的基因子串是以100结尾的。对于这样的交叉繁殖，要想得到更好的结果只能通过基因突变来实现，而基因突变可能需要经过很多年或者好多代人。在未来

的繁殖中，当基因串在一个随机的位置分裂组合时，也可以更好的结果。另一种可能的解决方案就是互换父母基因中的随即选择位。

遗传规则通常都是带来一个好的结果，但有时需要很多年的时间。这个结果最接近于全局最大值，但不是最好的。在平坦的函数中，分级方法收敛得越快，得到的结果也越好。而遗传规则就相对较慢，但它更加稳健。

名词解释

后向传输：多层神经网络的训练技术。

双极性神经单元：神经单元的输出信号值处于 -1~1 之间。

前馈网络：没有反馈的网络。

感知器：具有固定门限值神经单元的网络。

循环网络：具有反馈的网络。

监督学习：理想输出值已知的学习过程。

单极性神经单元：神经单元的输出信号值处于 0~1 之间。

无监督学习：理想输出值未知的学习过程。

参考文献

[1] Fahlman, S. E. 1988. Faster-leaining variations on backpropagation: an empirical study. *Proceedings of the Connectionist Models Summer School*, eds. Touretzky, D. Hinton, G., and Sejnowski, T. Morgan Kaufmann, San Mateo, CA.

[2] Fahlman, S. E. and Lebiere, C. 1990. The cascade correlation learning architecture. *Adv. Ner. Inf. Proc. Syst.*, 2, D. S. Tourezky, ed. pp. 524-532. Morgan, Kaufmann, Los Altos, CA.

[3] Goldberg, D. E. 1989. *Genetic Algorithm in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA.

[4] Grossberg, S. 1969. Embedding fields: a theory of learning with physiological implications. *Journal of Mathematical Psychology* 6: 209~239.

[5] Hebb, D. O. 1949. *The Organization of Behivior, a Neuropsychological Theory*. John Wiley, New York.

[6] Hecht-Nielsen, R. 1987. Counterpropagation networks. *Appl. Opt.* 26 (23): 4979~4984.

[7] Hecht-Nielsen, R. 1988. Applications of counterpropagation networks. *Neural Networks* 1: 131~139.

[8] Holland, J. H. 1975. *Adaptation on Natural and Artificial Systems*. University. Of Michigan Press, Ann Arbor, MI.

[9] Hopfield, J. J. 1982. Neural Networks and physical systems with emergent collective computation abilities. *Proceedings of the National Academy of Science* 79: 2554-2558.

[10] Hopfield, J. J. 1984. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Science* 81: 3088-3092.

[11] Kohonen, T. 1988. The neural phonetic typerater. *IEEE Computer* 27 (3): 11~22.

[12] Kohonen, T. 1990. The self-organized map. *Proc. IEEE* 78 (9): 1464~1480.

[13] Kosko, B. 1987. Adaptive bidirectional associative memories. *App. Opt.* 26: 4947 ~ 4959.

[14] Kosko, B. 1988. Bidirectional associative memories. *IEEE Trans. Sys. Man, Cyb.* 18: 49 ~ 60.

[15] McCulloch, W. S. and Pitts, W. H. 1943. A logical calculus of the ideas imminent in nervous activity. *Bull. Math. Biophy.* 5: 115 ~ 133.

[16] Minsky, M. and Papert, S. 1969. *Perceptrons.* MIT Press, Cambridge, MA.

[17] Nilsson, N. J. 1965. *Learning Machines: Foundations of Trainable Pattern Classifiers.* McGraw-Hill, New York.

[18] Nguyen, D. and Widrow, B. 1990. Improving the learning speed of 2-layer neural networks, by choosing initial values of the adaptive weights. *Proceedings of the International Joint Conference on Neural Networks* (San Diego), CA, June.

[19] Pao, Y. H. 1989. Adaptive Pattern Recognition and *Neural Networks.* Addison-Wesley, Reading, MA.

[20] Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization on the brain. *Psych. Rev.* 65: 386 ~ 408.

[21] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning internal representation by error propagation. *Parallel Distributed Processing.* Vol. 1, pp. 318 ~ 362. MIT Press, Cambridge, MA.

[22] Sejnowski, T. J. and Rosenberg, C. R. 1987. Parallel networks that learn to pronounce English text. *Complex Systems* 1: 145 ~ 168.

[23] Specht, D. F. 1990. Probalistic neural networks. *Neural Networks* 3: 109 ~ 118.

[24] Specht, D. F. 1992. General regression neural network. *IEEE Trans. Neural Networks* 2: 568 ~ 576.

[25] Wasserman, P. D. 1989. *Neural Computing Theory and Practice.* Van Nostrand Reinhold, New, York.

[26] Werbos, P. 1974. Beyond regression: new tools for prediction and analysis in behavioral sciences. Ph. D. diss., Harvard University.

[27] Widrow, B. and Hoff, M. E., 1960. Adaptive switching circuits. 1960 IRE Western Electric Show and Convention Record, Part 4 (Aug. 23): 96 ~ 104.

[28] Widrow, B. 1962. Generalization and information storage in networks of adaline Neurons. In *Self-organizing Systems*, Jovitz, M. C., Jacobi, G. T. and Goldstein, G. eds. Pp. 435 ~ 461. Sparten Books, Washington, D. C.

[29] Wilamowski, M. and Torvik, L. 1993. Modification of gradient computation in the backpropagation algorithm. *ANNIE' 93—Artifical Neural Networks in Engineering.* November 14-17, 1993, St. Louis, Missou.; also in Dagli, C. H. ed. 1993. *Intelligent Engineering Systems Through Artificial Neural Networks* Vol. 3, pp. 175 ~ 180. ASME Press, New York.

[30] Zadeh, L. A. 1965. Fuzzy sets. *Information and Control* 8: 338 ~ 353.

[31] Zurada, J. M. 1992. *Introduction to Artificial Neural Systems.* West Publishing Company, St. Paul. MN.

第19章　机器视觉

David A. Kosiba　Rangachar Kasturi

19.1　引言

机器视觉也称为“计算机视觉”，它是一门科学，利用视觉图像原理来对我们周围世界中的物理对象进行清晰、准确地描述。机器视觉根据物理对象的几何属性和构成技术来得到它的尺寸和抽象性，从而用来评估图像特征，同时将特征尺寸与空间对象的几何学相互关联，最后形成几何信息。上述这个过程通常称为“图像理解”。

机器视觉系统的目标就是根据图像或图像序列来产生一个真实的空间模型。由于图像是对三维空间的二维映射，因此其信息就不能直接使用，必须进行重新复原。这种复原技术需要一个反向的多对一映射。但是，为了再次利用图像信息，就需要了解对象在环境和投影几何学方面的知识。

在机器视觉系统每一个阶段的决策过程中，必须深刻理解应用目标的相关知识。由于机器视觉系统的重点是使每一个阶段尽可能实现自动化操作，因此，掌握相关的知识是实现机器视觉系统的前提。系统中使用到的知识包括特征模型、图像信息、对象以及对象之间的相关性。如果没有准确的相关知识，机器视觉系统的应用范围就会变得很狭隘。为了进一步提高灵活性和耐用性，系统描述的相关知识就必须详细准确，而且可以直接使用。不过，机器视觉系统设计师有时也会使用一些隐藏的相关知识，如同使用准确描述的相关知识一样。事实上，一个机器视觉系统的效能和效率是由系统所使用知识的质量决定的，系统中的难题通常只能通过合适的知识来源以及系统中相应的应用机制才能解决。

1. 与其他领域的关系

机器视觉与很多其他学科密切相关，它融合了各种各样的技术，这些技术很多来自于非常完善的学科领域，例如物理学、数学和心理学。从图像中复原信息的技术也是从很多领域中开发出来的。在本节中，我们将主要描述一些与其密切相关的领域。

图像处理技术通常是指将图像从一种格式转换成其他格式，而信息复原过程则留给用户自己进行。图像处理领域中包含的主题包括图像增强、图像压缩以及模糊图像校正（Gonzalez 和 Woods，1982）。换句话说，机器视觉系统将图像作



为输入信号，同时输出其他类型的信号。例如，图像中对象轮廓线的描述或者一系列图像中对象的运动轨迹。因此，机器视觉系统的重点是自动复原信息，尽可能减少人为参与。图像处理技术在机器视觉系统的开始阶段是非常有用的，可以用来增强有效信息，并抑制干扰。

计算机制图技术是指根据几何图元（例如线、圆和自由面）来生成图像。计算机制图技术在可视化和虚拟现实系统（Foley 等，1990）中占据了重要地位；而机器视觉学科则是一个相反的过程，即根据图像来评估几何图元和其他特征。因此，计算机制图是指图像的合成，而机器视觉则是指图像的分析。但是，这两个领域正在逐渐靠拢。机器视觉学科使用了计算机制图技术中的曲线图和表面描述方式以及其他技术；而计算机制图技术也使用了机器视觉学科中的很多技术来进行计算机建模，以生成现实的图像。可视化和虚拟现实系统使得这两个领域变得越来越接近了。

模式识别技术用来对数字形式的数据和符号形式的数据进行分类，很多用于模式分类的统计技术和合成技术已经被开发出来了。模式识别技术在机器视觉系统的对象识别过程中扮演着重要角色。事实上，很多可视化应用技术在很大程度上依赖于模式识别。机器视觉系统中的对象识别过程通常需要很多其他技术。关于模式识别技术的详细讨论，可以参考 Duda 和 Hart 在 1973 年的相关著作。

人工智能技术是指与智能系统设计和计算机智能（Winston，1992；Tanimoto，1995）研究相关的技术。人工智能通常用来在图像经过处理得到特征参数后，通过计算场景内容的符号表达式来对其进行分析。人工智能可以划分为 3 个阶段：感知、识别和处理。感知阶段将现实对象中的信号转换成符号；识别阶段负责符号的处理；处理阶段则负责将符号转换成能影响现实对象变化的信号。人工智能的很多技术在机器视觉学科的各个方面扮演着重要角色。事实上，机器视觉学科经常被认为是人工智能学科中的一个分支。与人工智能直接相关的就是神经网络学科。神经网络的设计和分析在近十年内已经发展成了一个非常活跃的领域（Kosko，1992），而且神经网络已经越来越多地被用来解决机器视觉系统中的很多问题。

精神物理学和认知科学领域研究人员共同长期研究人类的视觉（Marr，1982）。机器视觉学科中的很多技术都和人类视觉有关。相比设计机器视觉系统，很多计算机视觉研究人员对根据人类视觉进行计算机建模更感兴趣。因此，机器视觉系统中使用的很多技术与精神物理学中的很多技术具有很强的相似性。

2. 视觉学科的基础

类似于其他正处于发展中的学科，机器视觉学科也是基于一定的基础学科和技术的。为了成功开发机器视觉系统，我们必须从初始的图像形成到最后的图像合成等角度来彻底理解系统的各个方面。下面列出了机器视觉学科的各个要点和

主题：

1）图像形成；

2）分割；

3）特征提取和匹配；

4）三维对象识别；

5）动态视觉。

在上面的每一步中，都有很多因素会影响到特定系统开发过程中机制和技术的选择。系统设计师必须对系统设计的要点和各种技术和机制之间的权衡有一个很好的把握。接下来，我们将主要讨论上述这些主题。在接下来的整个章节中，我们将不会引用所有的原著，因为这些原著太多了。但是，读者可以参考本章后面给出的参考文献。如果读者想了解关于这些主题和其他主题更详细的讨论，我们建议读者参考备注部分的相关文献。

19.2 成像过程

当一个传感器记录下了接收到的放射光线，并将其作为一个二维函数时，一个图像就形成了。图像中的亮度值或强度值代表了不同的物理实体。例如，一个摄像机中得到的图像，其亮度值就代表了空间中对象各个面的反射系数；在热成像中，其亮度值就代表了空间中相应区域的温度；在距离成像中，其亮度值就代表了镜头与空间中各个点之间的距离。同一个空间中的多个图像通常采用不同类型的传感器来捕捉，从而可以获得更好的、更可靠的空间数据。选择合适的成像系统对于实际机器视觉系统的设计至关重要，接下来本节将详细描述成像的原理。

1. 成像几何学

图 19-1 给出了一个简单的摄像中心模型示例。坐标系统中 xy 平面与成像平面相互平行，而 z 轴穿过镜头中心，并与成像面相交于距离 f 处，f 为摄像机的焦距。空间点（x，y，z）的图像在成像面上形成了一个点（x'，y'），其中，

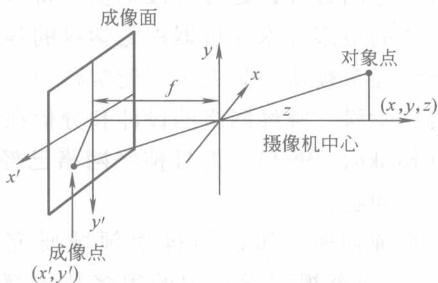


图 19-1 摄像中心坐标系统成像模型

$$x' = x\frac{f}{z},\ y' = y\frac{f}{z} \tag{19-1}$$

这就是成像系统中的立体投影公式。

在成像过程中，摄像机可能具有多个角度，例如平移镜头、摇拍镜头和倾

斜镜头。同样，多个摄像机也可以从不同的点来拍摄相同的场景。在本例中，可以方便地使用空间坐标系统，而场景空间和摄像机的坐标就是根据该坐标系统来定义的，但其成像公式就会变得非常复杂。因此，建议读者参考本章最后的参考文献中关于成像几何学更详细的描述，其中就包含摄像校准方面的描述。

2. 图像亮度

尽管成像几何学系统地定义了场景空间坐标与图像坐标之间的关系，但是图像中每个点的亮度或强度不仅取决于成像几何学，也取决于很多其他因素，包括景物亮度、反射系数特性和场景中对象的表面方位以及成像传感器的辐射度属性。

表面的反射系数特性是通过它的“双向反射分布函数（Bidirectional Reflectance Distribution Function，BRDF）”来描述的。BRDF 是指观察者方向的场景辐射强度与定向光源单位时间内投射到单位面积上的辐射能量的比值。BRDF 说明了当从另一个方向照射而从一个给定的方向观察时，物体表面将会产生多大的亮度。例如，对于一个朗伯（Lambert）表面来说，如果照射光源是一个很远的点，那么 BRDF 值就是一个常量，因此物体表面在各个方向上来看亮度都是相同的；而对于一个镜面（类似镜面）来说，BRDF 就是一个冲激函数，该函数与反射规律有关。

物理对象表面上某个点的辐射强度取决于其表面的反射系数特性以及照射光源的亮度和方向。例如，对于一个被点光源照射的朗伯（Lambert）表面来说，辐射强度就和表面法线与照射方向之间夹角的余弦成正比。这种表面方位和亮度之间的关系在“反射系数图”中得到了体现。在指定表面和照射强度的反射系数图中，相同亮度的等值线是通过一个表面方位函数来表示的，该函数由斜率空间坐标（p，q）表示，其中 p 和 q 分别表示 x 方向和 y 方向的表面斜率。图 19-2 给出了被点光源照射的朗伯（Lambert）表面的反射系数图。在该图中，最亮的点对应了表面方位，这样正位点都处于光源方向上。由于图像亮度与场景辐射强度成正比，因此某一点的图像亮度就与相应场景点的表面方位直接相关。“描影法”就是利用这种关系来复原三维目标的形状。“光学立体系统”也利用了相同的原理，它根据从不同方向照射的场景中得到的多个图像来复原目标的形状。

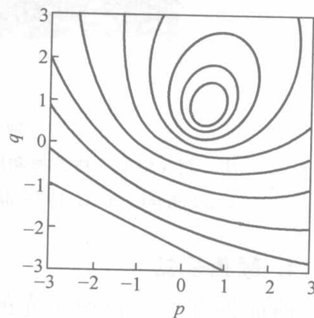


图 19-2 被点光源照射的朗伯（Lambert）表面的反射系数图

注：照射方向为 $(p,q)=(0.2,0.4)$

3. 采样与量化

在计算机中，连续函数是无法表示的。成像系统和计算机之间的接口必须对连续图像函数进行有限采样，并为每个采样值指定一个有限位码，这个过程称为“采样与量化”。每个图像采样值都是一个“像素”。通常，图像的采样都是一格小方块，这样整个图像中所有像素之间的水平距离和垂直距离都是相同的。每个像素在计算机中都是一个整数，并表示成一个二维的数组。通常，一个像素会被表示成一个无符号的 8bit 整数，范围为（0~255）；其中 0 对应的是黑色，255 对应的是白色，灰色分布在中间。

很多摄像机得到的都是模拟图像，该模拟图像稍后会被采样和量化，从而转换成数字图像。采样率决定了数字图像拥有多少像素（图像分辨率），而量化过程决定了在表示每个采样点的亮度值时使用的灰阶数量。如图 19-3a 和图 19-3b 所示，在不同的采样率和量化亮度下的图像看上去有很大区别。在大多数机器视觉应用中，采样率和量化亮度都是根据摄像机和图像获取硬件的有限可用选择预先确定的。但是，在很多应用中，了解采样率和量化亮度的影响是很重要的。

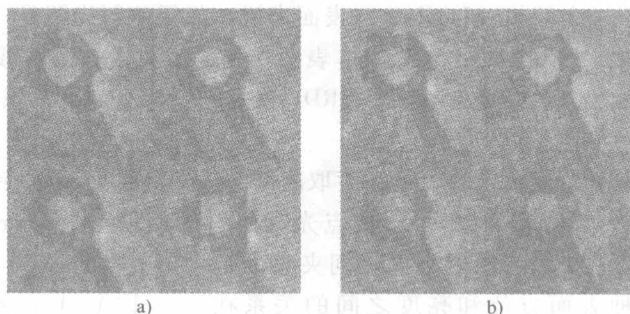


图 19-3 采样图像

a）各种空间分辨率 b）不同数量的量化亮度

注：图 a 中的方块结构和图 b 中出现的模糊轮廓（来源：Jain，R.，Kasturi，R.，and Schunk，B. G. 1995. *Machine Vision*. McGraw-Hill，NewYork. 引用已经过许可）

4. 颜色视觉

前面的部分只讨论了光的强度，但是，光是由不同波长的光谱组成的。图像中可以包含很多不同波长的采样点，从而可以形成彩色的图像。对颜色的感知取决于：场景表面的光谱反射系数；照射光源的光谱结构；成像系统中的光谱灵敏度。在人体系统中，颜色感知主要得益于视网膜中上百万个神经化学感光单元（杆状细胞和视锥细胞）在光谱灵敏度上的差别。杆状细胞主要负责单色视觉的感知，而视锥细胞主要负责颜色的感知。人体视觉系统由 3 种不同类型的视锥细胞构成，每一种都具有不同的感光灵敏度。每个视锥细胞的总体灵敏度通常由下面的积分来模拟：

$$R_i(\lambda)=\int f(\lambda)r(\lambda)h_i(\lambda)d\lambda \quad i=1,2,3 \tag{19-2}$$

式中，λ 是指入射到感知单元的光波波长；$f(\lambda)$ 是指照射光源的光谱构成；$r(\lambda)$ 是指反射面的光谱反射函数；$h_i(\lambda)$ 是指第 i 类感知单元（视锥细胞）的光谱灵敏度。

在机器视觉系统中，彩色图像通常是通过3种不同类型的传感器响应来表示的（例如，主要颜色为红、绿和蓝），其中每一种传感器都具有不同光谱灵敏度。这3种主色的选择决定了通过加权组合可以实现的颜色范围。但是，传感器的灵敏度不必局限于可见光。例如，传感器也可以轻松地对电磁频谱中红外、紫外或X射线波段的波长产生响应，其中任何一种情况下，成像系统都会产生多个图像，称为“波道”，每个传感器对应一个波道。如果3个传感器的输出信号都被使用，那么该过程对图像处理和存储器的要求就提高了3倍。但是，这种要求也可以通过增加维数来进行弥补，这种技术目前已经得到应用了。

5. 距离成像

虽然三维信息可以通过图像结构（如阴影、纹理和运动）从二维图像中提取，不过这个过程在距离成像中就变得更加容易了。在距离成像过程中，图像中的每一个像素值都是目标对应点与传感器之间距离的函数，最后形成的图像称为“距离图像”或“深度图”。图19-4给出了一个咖啡杯的简单距离图像示例。在此，我们将主要讨论两种常见的距离成像技术，即“成像雷达”和“三角测量法”。

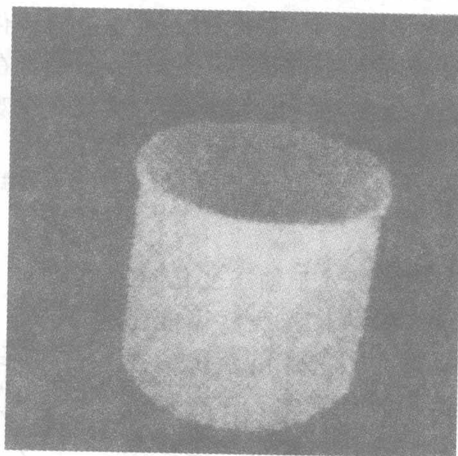


图19-4 咖啡杯的距离成像

（1）成像雷达

在飞行脉冲调制雷达系统中，目标的距离是通过观察发射电磁脉冲和接收电磁脉冲之间的时间差来计算的。距离信息也可以通过检测调幅波束中发射电磁波和接收电磁波之间的相位差得到，或者通过检测调频波束中相干混合发射-接收信号的差频得到。

（2）三角测量法

在基于三角测量法的动态距离成像系统中，光发射器和摄像机被放置在 z 轴上，两者之间的基线距离为 b，如图19-5所示。目标坐标（x,y,z）与图像坐标（x',y'）和发射夹角 θ 相关，公式为

$$[xyz]=\frac{b}{f\cot\theta-x'}[x'y'f] \qquad (19\text{-}3)$$

精确度决定了该三角测量系统的距离分辨力，其中夹角 θ 和水平位置 x' 可以通过精度来测量得到。动态三角测量系统采用了单点光源，该光源以光栅扫描方式穿过场景目标。

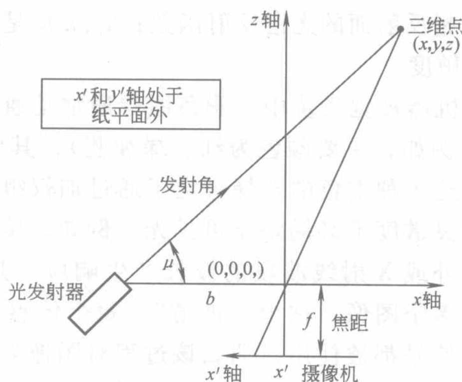


图 19-5　基于三角测量法的距离成像系统（来源：Best, P. J. 1988. Active, optical range imaging sensor. Machine Vision and Applications 1 (2): 127-152）

6. 结构化光源

采用结构化光源的成像系统，其场景中使用的光源是一个有名的几何光模型。在一个简单的点光源发射系统中（例如前面描述过的三角测量系统），某一时刻只有一个点光源进行二维形式的照射。每个点的深度是根据前面的公式来计算的，从而获得距离图像。在一个结构化光源系统中，每一个光面或二维光图像都会被应用到场景中去。摄像机被放置在距离发射光源一定位移的位置上，用来记录目标光表面上的光图像。光图像中观察到的扭曲现象是由目标表面的形状和方位决定的；在目标表面，图像会被反射，如图 19-6 所示。

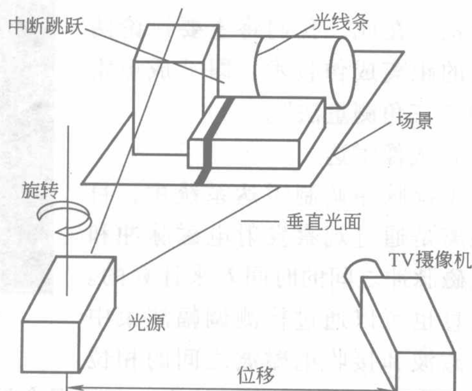


图 19-6　结构化光源系统（来源：Jorvis, R. A. 1983. A perspective on range finding techniques for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5 (2): 122-139. ©1983 IEEE.）

当场景处于动态变化时，依次发射单束光线就可能不太实际了，这是由系统的连续性本质决定的。在这种情况下，就必须使用一组多束光线（每一组都被惟一进行编码）。例如，采用二进制编码方案，只需要通过发射 $\log_2 N$ 个图像就可以得到一组完整的数据；其中，$N-1$ 是每个图像中光线的总条数。其他的结构化光源技术包括彩色

编码、网格编码、傅里叶变换域处理。任何结构化光源系统的主要缺点就是数据不能从目标点处得到，目标点对于光源和成像摄像机来说都是不可见的。

7. 有源视觉

大多数机器视觉系统依赖于固定硬件捕捉的数据，这些系统包括无源传感系统（如摄像机）和有源传感系统（如激光距离探测器）。在有源视觉系统中，数据捕捉过程的参数和特性是由场景描述系统动态控制的。有源视觉系统既可以使用有源传感器，也可以使用无源传感器。但是，在有源视觉系统中，传感器的状态参数（如焦点、孔径、聚散度和光照强度）必须得到有效控制，以获得有助于场景描述的数据。

19.3 分割

一个图像必须进行分析，并在进行更加抽象地描述和表示之前提取出相应的特征。选择合适的所谓低级操作对于成功描述更高级的场景是至关重要的。机器视觉系统首先要完成的操作就是将目标与背景进行区分。该操作（又称为“分割”）可以通过两种方式实现：基于边缘的方法，该方法可以定位在某些特性方面的不连续性；基于区域的方法，该方法可以根据某些相似性来对像素进行分组。

1. 基于边缘的分割方法

在基于边缘的分割过程中，通常采用目标的边界来对图像进行划分。处于目标边界上的点必须被标识出来，这些点称为“边缘点”；我们通常可以通过分析某个点的局部相邻性来查找这些点。根据定义，边缘点两侧的区域（例如目标和背景）具有明显不同的特性。因此，在边缘检测中，重点就是检测相邻点之间的不同特性。

（1）斜率

图像中的边缘是通过图像亮度中的明显局部变化来显示的，通常与图像亮度或图像亮度一阶导数的不连续性直接相关。因此，一个直接的边缘检测方法就是计算图像中每个像素的斜率。斜率是一个二维矢量的一阶导数，定义如下

$$G[f(x,y)] = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \dfrac{\mathrm{d}f}{\mathrm{d}x} \\ \dfrac{\mathrm{d}f}{\mathrm{d}y} \end{bmatrix} \tag{19-4}$$

式中，$f(x,y)$ 是指二维图像亮度函数。

根据一阶偏导数 G_x和 G_y，斜率的大小和方向可以表示如下

$$G(x,y) = \sqrt{G_x^2 + G_y^2} \quad \alpha(x,y) = \tan^{-1}\left(\frac{G_y}{G_x}\right) \tag{19-5}$$

只有那些斜率比预先指定门限值大的像素才能被认为是边缘像素。

在计算 x 方向和 y 方向的偏导数时，我们可以使用很多近似的方法。作为早期的边缘检测器之一，Robert 交叉算子是通过将对角线上像素的亮度差放进一个 2×2 的窗口来计算偏导数的。其他的近似方法，例如 Prewitt 算子和 Sobel 算子是通过一个 3×3 的窗口来计算偏导数的。有时候，计算各个指定方向边缘亮度的算子也会用于机器视觉系统中，这些定向算子的窗口通常超过 3×3 或者更大。而对于那些简单边缘检测算子不适用的应用来说，通常使用 Canny 边缘检测器，该检测器可以优化边缘位置的干扰压缩。图 19-7 给出了上述每个边缘算子中的窗口算子描述示例。

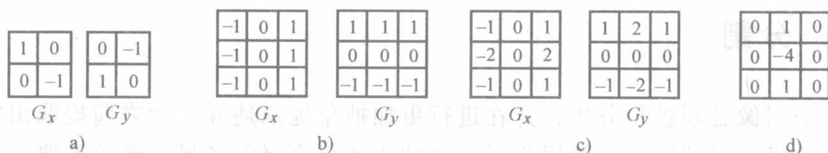


图 19-7　各种算子

a）Robert 算子　b）Prewitt 算子　c）Sobel 算子　d）拉普拉斯（Laplacian）算子

注：拉普拉斯（Laplacian）算子不是定向算子

（2）拉普拉斯算子

当使用基于斜率的边缘检测器时，相邻的像素通常会获得多个响应，这是由于边缘的过渡率决定的。在经过门限处理后，通常会得到很厚的边缘。但是，一个边缘只需要一个响应；换句话说就是，只需要最大的斜率，而且不是每一个响应值都高于指定的门限值。可以通过检测图像亮度函数中二阶导数的零值点来找到最大的斜率。根据二阶导数，拉普拉斯算子的定义如下

$$\nabla^2 f=\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2} \tag{19-6}$$

x 方向和 y 方向的二阶导数可以通过差分方程近似得到

$$\frac{\partial^2 f}{\partial x^2}=f(x+1,y)-2f(x,y)+f(x-1,y)$$

$$\frac{\partial^2 f}{\partial y^2}=f(x,y+1)-2f(x,y)+f(x,y-1) \tag{19-7}$$

通过合并上面的两个方程式，就可以在单个窗口算子中实现拉普拉斯（Laplacian）算子，如图 19-7 所示。

拉普拉斯算子的响应可以通过将图像与合适的窗口算子进行卷积来计算。不过，由于拉普拉斯算子不是精确的二阶导数，因此，它对线条、线条终端以及干扰像素的响应比对边沿的响应强烈，从而每个边沿都会产生两个响应：一个正响

应，一个负响应。因此，两个响应之间的零过渡区域就可以用来区分边沿了。

（3）高斯拉普拉斯算子

根据图像亮度函数二阶导数的零值点来检测边缘点的过程会出现干扰，因此，通常在进行边缘检测之前需要将干扰过滤掉。为了解决这个问题，高斯拉普拉斯（Laplacian of Gaussian，LoG）算子合并了高斯滤除算法和拉普拉斯边缘检测算子。LoG 算子的输出值 $h(x,\ y)$ 可以通过卷积操作得到，如下所示：

$$h(x,y)=\nabla^2(g(x,y)f(x,y)) \tag{19-8}$$

式中，

$$g(x,y)=\mathrm{e}^{-\left[\frac{(x^2+y^2)}{2\sigma^2}\right]} \tag{19-9}$$

$g(x,\ y)$ 是高斯滤波器的方程式（忽略常量因子）。利用卷积的导数规则，我们可以得到等价的表达式，如下所示：

$$h(x,y)=[\nabla^2 g(x,y)]f(x,y) \tag{19-10}$$

式中，

$$\nabla^2 g(x,y)=\left(\frac{x^2+y^2-2\sigma^2}{\sigma^4}\right)\mathrm{e}^{-\left[\frac{(x^2+y^2)}{2\sigma^2}\right]} \tag{19-11}$$

$\nabla^2 g(x,y)$ 通常称为“Mexican Hat”算子，因为它只有在进行分割时才出现，如图 19-8 所示。

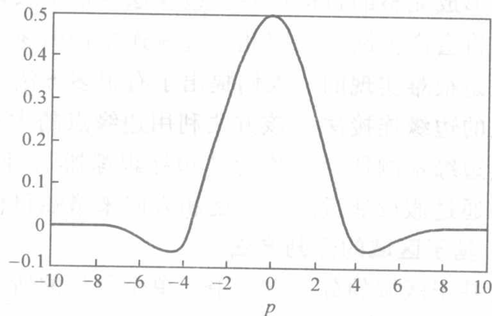


图 19-8　高斯拉普拉斯函数的周期

上面分析的结果显示，高斯拉普拉斯算子可以通过两种方式得到：将图像与高斯平滑滤波器进行卷积，并计算结果的拉普拉斯算子；将图像与高斯拉普拉斯滤波器进行卷积。零值点仍然必须检测出来，以便对边缘进行定位；而且类似于拉普拉斯算子，只有那些一阶导数高于指定门限值的边缘点才会被标识出来。

当采用高斯拉普拉斯方法时，边缘检测的准确性（比例）主要取决于高斯滤波器的使用广泛性。任何图像中，在确定真实边缘时需要利用分级空间法来融合不同亮度的算子信息。具有一定准确性的边缘检测对应了一个合适的亮度，而且边缘的准确位置可以通过利用低 σ 值来追踪得到。

（4）曲面拟合

由于数字图像实际上是一个二维空间变量的采样连续函数，因此另一种用来检测边缘的可行方法就是近似和重建图像的基本连续空间函数，这个过程称为“曲面拟合”。图像中某个点附近的亮度值可以用来获取基本连续亮度表面，该

表面可以用来作为最佳近似值，利用该近似值可以计算该点附近的特性。连续亮度函数可以表示成

$$z=f(x,y) \tag{19-12}$$

如果图像可以描述一个表面，那么式（19-12）就可以准确地描述图像的特性。但是，如果一个图像包含了很多个表面，这种情况下，上面的公式就只能满足局部表面点的应用。因此，这种方法利用了每个像素附近的亮度值来近似该点的局部表面。小平面模型就是这种方法的典型例子。在小平面模型中，某个点附近的区域就是通过立方曲面来近似的，该曲面是该区域的最佳近似，公式如下：

$$f(r,c)=k_1+k_2r+k_3c+k_4r^2+k_5rc+k_6c^2+k_7r^3+k_8r^2c+k_9rc^2+k_{10}c^3 \tag{19-13}$$

式中，r 和 c 是指附近区域中心点的局部坐标。

根据这种近似方法，我们可以计算二阶偏导数，并检测零值点，还可以找到边缘点。

（5）边缘连接

边缘像素的检测只是分割任务中的一部分工作，边缘检测器的输出必须连接起来以形成完整的目标边界。边缘像素很少来自于邻近的边界，遗失的边缘点在边界中将会产生缺口。因此，边缘连接在图像分割中就变得异常重要了，但是这个过程是很难实现的。人们提出了有很多方法来提高边缘检测器的性能，包括基于张驰的边缘连接法（该方法利用边缘点的大小和方向来定位附近的其他边缘）和连续边缘检测器（也称为“边缘跟踪器”，该检测器可以发现亮度较大的边缘点，并通过假设邻近边缘亮度和方向来增强目标边界）。

2. 基于区域的分割方法

在基于区域的分割方法中，单个目标的所有像素都会被聚集在一起并被标识出来，以表示它们属于相同的目标。像素点的聚集过程是基于一定规范的，该规范可以将属于相同目标的像素与其他像素区分开。具有类似特性的像素点就会被看成一样的像素点，并划分成一组互连的点，称为“区域”。这些像素点通常属于单个目标或某个目标的一部分。利用这种基于区域的分割方法来分割图像时，有很多种技术可供选择。

（1）区域形成

分割过程通常是从简单区域形成开始的。在该步骤中，利用图像的本质特性来形成初始的区域，从图像亮度矩形图中得到的门限值通常用来实现这种初始的分组过程。一般来说，每个图像都包含多个区域，每一个区域又具有不同的本质特性。这种情况下，图像亮度矩形图会出现几个峰值，每一个峰值都可能对应一个或多个区域，这样这些峰值就可以用来选择多个门限值。经过门限处理之后，一个互连结构的算法就可以用来查找初始区域了。

门限处理技术通常会产生过多的区域。由于这些区域在形成时是基于一阶特性的，因此得到的区域通常过分简单化，而且无法对应完整的目标。通过门限处理得到的区域可以看做是分割过程的第一步。在基于矩形图的初始分割过程之后，可以使用更多更成熟的技术来对分割过程进行精炼。

（2）分裂与合并

基于亮度的初始分割其自动精炼过程通常是通过分裂与合并的组合操作来实现的。分裂与合并操作可以通过分裂一个区域（该区域中包含多个目标的图像）或者合并相邻的区域（实际上属于相同的目标）来消除模糊边界和虚假区域。如果一个区域的某些特性不是一致的，那么该区域就应该被分裂。基于分裂方法的分割过程是从较大的区域开始的。在很多情况下，整个图像也可能作为起始区域。在对一个区域进行分裂之前，必须确定以下几点：一是确定整个区域上特性是在“何时”出现不一致的；二是确定“如何”对区域进行分裂，以便最后得到的每一个子区域的特性是一致的。上面这两点通常不是很容易确定的。在某些应用中，亮度值的变化成为了一致性的衡量标准。

比确定属性的一致性更难的是决定从“哪里”分裂一个区域。基于属性值来进行区域分裂是非常困难的。一种方法就是当试图通过确定最佳边界来划分区域时，在区域内假设各种边缘值。分裂区域最简单的方案是将区域划分成一组固定数量的相同区域，这种方法称为“常规分解法”。例如，在“四叉树”方法中，如果一个区域被认为是不一致的，那么该区域在每一步中将会被分裂成4个相同大小的象限区域。

人们提出了有很多种方法来判决区域的相似性。广泛使用的方法要么基于区域特性，要么基于区域之间边缘的薄弱点。下面是两种基于区域特性来判决相邻区域相似性的常见方法：比较它们的平均亮度；假设它们的亮度值来自于一个著名的概率分布。在第一种方法中，如果两个区域之间的平均亮度差没有超过预先定义的值，那么这两个区域就会被认为是相似的，并将成为合并的对象。这种方法的改进型是利用曲面拟合来确定两个区域是否可以近似成一个区域。在第二种方法中，相邻的区域能否合并取决于两个区域是否具有相同的亮度值统计分布概率。这种方法采用假设测试法来判决相邻区域的相似性。

合并的另一种方式就是当两个区域之间的边界很模糊时可以合并这两个区域，模糊边界是指两边的亮度差小于给定的门限值。这种方法可以通过假设亮度特性和共同边界的长度来消除两个相邻区域之间的模糊边界。如果共同的边界是模糊的而且最后合并区域的边界也不明显的话，那么共同的边界就会消失。

分裂与合并操作也可以一起进行。当基于门限处理的预分割完成后，一组连续的分裂与合并操作就可以根据区域的特性进行了。这种方案已经被用于复杂场景的分割了。关于控制分裂与合并操作的相关领域知识在本章中也将进行介绍。

3. 其他分割方法

在前面的讨论中，我们主要集中在图像的亮度上，而基于颜色、纹理和运动的分割技术已经被开发出来了，而且基于光谱模式分类的分割技术已经广泛应用于遥感系统中了。

19.4 特征提取和匹配

分割后的图像通常是由一个压缩形式的图像来进行描述的，该压缩图像可以进一步进行特征提取。我们选择的图像表示方案必须与目标识别和描述的方法相匹配。目标识别任务要求将图像中的目标描述与已知的目标模型相匹配起来。反过来，模型也可以使用某些描述性的特征以及它们之间的关系。匹配过程同样在图像复原信息的其他方面扮演着重要角色。接下来，我们将讨论机器视觉系统中常见的表示和描述方案，同时还将介绍特征提取和匹配技术。

1. 表示法和描述法

符号信息的表示和描述可以通过很多种方式实现。一种方式就是根据目标的边界曲线来表示。常见的边界表示法包括循环码、多边形法、一维标记法以及利用要点来表示目标的方法。另一种方式就是获取基于区域的形状描述符，例如拓扑或纹理描述符。表示法和形状描述方案通常会有选择性，以便旋转、转换和比例变化时描述符不会发生变化。

（1）循环码

最早利用双向码来表示一个边界的方法之一就是“循环码”。目标边界会在一个合适的亮度下进行重新采样，沿边界的一列整齐的点就通过一串方向码来表示，如图19-9a所示。通常为了将所有的信息都保留在边界上，重新采样过程会绕开边界。但是，重新采样过程消除了由于干扰导致的弱波动。循环码包含一些具有吸引力的特征；也就是说，一个目标以45°旋转是很容易实现的；循环码的导数在旋转时是不变的，差分码可以通过计算循环码的一阶微分获得。而一个区域的其他特性（如面积、夹角）也可以直接从循环码中计算得到。这种表示方法的局限性在于表示某个点的切线方向时受到了限制。尽管偶尔会遇到具有很多方向的代码，但是我们最常见的是8-方向的循环码。

（2）多边形法

边界的多边形近似法已经被广泛研究过，而且还开发出了很多种方法。多边形拟合可以减少近似曲线和原始曲线之间的误差。在“迭代终点”方案中，第一步就是在边界上两个最远的点之间连接一条直线段，这样就可以测量出分块与曲线上各个点之间的垂直距离。如果任何一个距离都大于选定的门限值，那么该直线段就会被两个直线段取代，其中任意一个直线段的终点到曲线点之间的距离

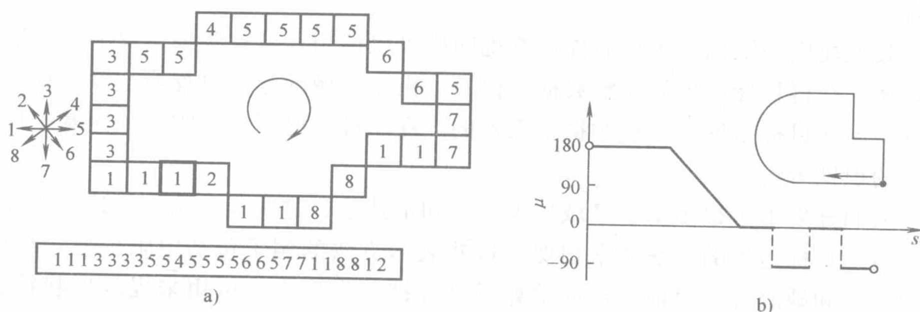


图 19-9　目标表示法
a）目标及其链式码　b）简单几何目标及其一维（s，θ）标记

都是最大的。另一种方式中，拟合直线会强制穿过每个数据点附近一定的范围。直线段从第一个点处开始出发，当直线段进一步延伸时，会导致直线段超出点的半径范围，这样一个新的直线段就形成了。我们还可以使用极小极大法，在这种方法中，直线段的近似是有选择性的，以便使数据点与近似直线段之间的最大距离最小化。

除了多边形近似法外，更高阶的曲线拟合与样条拟合法也可以用于要求更加准确的近似过程中。这两种方法在计算量上比多边形法更加复杂，因此很难用于实际应用中。

（3）一维标记法

在很多应用中，曲线中切线的斜率可以用来表示曲线，斜率由 θ 定义，其中 θ 是 s 的函数，s 是任意出发点沿曲线的方位。这些函数的曲线具有一些很有趣和很有用的特性。$s-\theta$ 曲线中的水平线代表了直线，而与水平直线的夹角代表了圆弧，该圆弧的半径与直线的斜率成正比。一个 $s-\theta$ 曲线也可以看做是一个周期函数，其周期由曲线的周长给定。因此，就可以使用傅里叶变换技术了。其他函数也可以作为形状标记，如从曲线内部任意出发点到曲线的距离可以表示成一个与水平夹角相关的函数。如图 19-9b 给出了一个简单的几何目标及其一维标记示例。

（4）边界描述符

由目标边界表示的目标描述符可以通过前面描述的表示法生成。简单的描述符（如周长、长度、长轴的方位、形状数量和离心率）都可以很容易地从边界数据中计算得到。边界上的一组整齐的点可以当作一个一维的复函数 $x_k + iy_k$，这些点具有二维坐标（x_k，y_k），其中 $k=1$，…，N，N 是边界上点的总数目。该复函数的离散傅里叶变换加权系数也可以作为一个形状描

述符。

其他使用一维标记函数的描述符也同样可以使用，但这些描述符的主要缺点就是无法获得目标边界的完整数据。但是，由于图像的分割过程或合并过程存在问题，因此目标边界的完整数据是无效的。在这种情况下，就需要采用基于局部信息的识别策略。

人们开发出了很多基于区域的表示法和描述法，这些表示法和描述法与那些用来表示目标边界的方法非常相似。这些表示方法的例子包括中轴变换、线框、凸起外壳和缺陷等。目前，运动学算子已经被开发出来，可用来提取形状特征，并生成目标形状的有效描述。拓扑描述符（如欧拉数）同样可以用来描述形状。小平面模型甚至可以生成图像的拓扑原始图，并通过 7 个描述性标签来命名部分拓扑图，这些标签包括凹陷点、峰值点、棱线等。这些描述符在目标的匹配过程中非常有用，基于区域的表示法和描述法对于区域中目标属性的描述也非常有用。例如，纹理对于目标的识别就非常重要。关于这些主题的完整讨论，读者可以参考本章最后的参考文献。

2. 特征提取

如果一个即将被识别的目标具有惟一的区别性特征，那么我们就会开发专用算法来提取这样的特征。目标匹配和识别的重点包括夹角、高曲率区域、拐点或曲线上曲率不连续的其他位置。在基于区域的匹配方法中，识别那些容易区分和识别的像素是很重要的。因此，人们提出了很多方法来检测曲线中的“要点”和区域中“令人感兴趣”的点。

（1）关键点

曲线中关键点（又称为“要点”，如夹角和拐点）的查找对于接下来的目标匹配和识别非常重要。大多数关键点的查找规则都会将局部曲率最大值点标识成要点。一种方法就是通过分析弧线中的曲线导数来查找曲线上的要点。被标识出来的点要么就是关键点，要么就属于平滑区间或干扰区间。这些标识方法主要与曲线是否与弧线有一定距离有关（或者靠近弧线，或者远离弧线）。另一种方法就是利用数学表达式来直接计算并标识轮廓线的曲率值。

（2）“令人感兴趣”的点

用于匹配两个图像的点必须是那些容易识别和匹配的点，这些点称为“令人感兴趣”的点。显然，那些统一区域和边缘中的点就不适合用来匹配。“兴趣”算子可以发现具有高方差的图像区域。在有些应用中（如根据运动来获取立体感和结构的应用），图像必须具有足够多的“兴趣”区域来协助匹配过程。

一个常用的“兴趣”算子——Moravec 算子，采用定向方差作为“兴趣”点的衡量标准。如果一个点具有局部最大的最小定向方差和，那么该点就会被认为是“兴趣”点。一个点的局部定向方差计算公式如下：

$$I_1 = \sum_{(x,y)\in s}[f(x,y) - f(x,y+1)]^2$$

$$I_2 = \sum_{(x,y)\in s}[f(x,y) - f(x+1,y)]^2$$

$$I_3 = \sum_{(x,y)\in s}[f(x,y) - f(x+1,y+1)]^2$$

$$I_4 = \sum_{(x,y)\in s}[f(x,y) - f(x+1,y-1)]^2 \quad (19\text{-}14)$$

式中，s 表示当前点的附近区域，其典型大小范围为 5×5 像素至 11×11 像素。该点的“兴趣值”如下所示：

$$I(x,y) = \min(I_1, I_2, I_3, I_4) \quad (19\text{-}15)$$

如果“兴趣值”$I(x,y)$ 是局部最大值，那么该特征点就会被选作“兴趣”点。该过程可以省去简单边缘点的检测，因为在边缘方向上的点没有方差。另外，如果一个点的局部最大最小定向方差和高于预选设定的门限值，那么该点就可以看做是一个良好的“兴趣”点。Moravec“兴趣”算子在立体匹配中已经得到了广泛应用。

3. 匹配

匹配过程在机器视觉系统的各个阶段中都扮演着非常重要的角色。目标识别过程中要求目标的描述与已知的目标模型相匹配。匹配的目标是：检测已知实体、目标或特征的存在；查找未知图像组成部分。实现这些匹配目标的第一个难点是目标直接匹配，目标直接匹配是指在图像中直接寻找特定的实体。一般来说，所有这些实体的位置都必须找到。在立体和运动结构应用中，各实体可以在一个图像中得到，而它们的位置由另一个图像确定。实现匹配目标的第二个难点是要求利用多个模型来匹配未知的实体，以确定哪一个模型匹配效果最佳。

（1）点式匹配

当对相同场景（例如立体图像或运动序列）中两个具有轻微区别的图中的点进行匹配时，“兴趣”点可以通过应用算子（如 Moravec“兴趣”算子）来进行检测。对应过程会充分考虑图像中被选点的局部结构，以便分配第二个图像中的初始可匹配候选点。但是，这个“对应过程”不是一个容易解决的问题，我们通常都会添加很多约束条件来简化该对应过程。例如，在立体匹配应用中，图像中的一个点到其他点的位移通常很小。因此，只有局部区域内的点才能作为匹配点。为了实现最后的对应关系，初始匹配可以通过计算每个候选点在整体结构上的相似性来进行精炼。在动态场景分析中，我们可以假设相邻点的运动不会发生很大的变化。为了实现两个图像中的最后对应关系，通常可以采用松弛法。图19-10给出了点式匹配的示例。

（2）图案匹配

在有些应用中，需要检测出图像中特定的图示或图标结构，这些结构称为“图案”。图案通常是由小型二维亮度函数来表示（通常小于 64 ×64 像素）。图案匹配是指图案跨越整个图像并检测出与图案最符合的图像位置的过程。通常用来确定匹配相似性的衡量标准称为“常规互相关性”，相关性加权系数计算公式如下所示：

来自图像1的感兴趣点 来自图像2的感兴趣点

图 19-10 点式匹配过程

$$M(x,y)=\frac{\sum\sum_{(u,v)\in R} g(u,v)f(x+u,y+v)}{\left[\sum\sum_{(u,v)\in R} f^2(x+u,y+v)\right]^{\frac{1}{2}}} \tag{19-16}$$

式中，$g(u,\ v)$ 为图案，$f(x,\ y)$ 为图像；R 为图案跨越的区域。M 是指点 $(x,\ y)$ 在 $g=cf$ 时的最大值（换句话说就是，图案的像素值和图像的像素值只需通过一个常数比例因子来进行区别）。

在经过门限处理之后，互相关性操作的结果即图案匹配的位置就可以得到了。图 19-11 给出了一个测试图像、一个图案和常规互相关性操作的结果。

objects in the *real*
object models.This
cognition effortlessl
ask for implementa
er we will discuss d
echniques that have
We will discuss diff

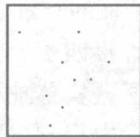
a)

e

b)



c)



d)

图 19-11

a）测试图像 b）字母“e”的图案（放大过） c）常规互相关性操作的结果

d）根据门限值得到的匹配位置（$T=240$）

（来源：Jain，R.，Kasturi，R.，and Schunk，B. G. 1995. *Machine Vision*. McGraw-Hill，New York. 引用已经过许可）

图案匹配技术的主要缺陷是其对目标的旋转和缩放比例过于敏感。为了匹配旋转和缩放的目标，必须构建独立的图案。在有些方法中，一个图案会被划分成多个子图案，然后匹配过程在这些子图案上进行。子图案之间的相互关系将会在最后的匹配阶段进行检验。

（3）Hough 变换

参数变换（如 Hough 变换）是直线和曲线识别中非常有用的方法。对于直线，参数空间变换是由直线的参数表达式来定义的，如下所示：

$$\rho = x\cos\theta + y\sin\theta \tag{19-17}$$

式中，(ρ, θ) 是参数空间中的变量，分别代表了初始点法线的长度和方向。

图像中每个点 (x, y) 都将转换成 (ρ, θ) 域的正弦曲线。但是，图像中所有点对应的正弦曲线都将相交于Hough域中的一个点上，这些点对应了图像面上的各个共线性像素。因此，直线上的像素可以很容易地被检测到。

Hough变换也可以通过不同的定义来识别其他类型的曲线。例如，圆上面的点可以通过搜索一个三维参数空间 (x_c, y_c, r) 来进行查找，(x_c, y_c, r) 中的前两个参数 (x_c, y_c) 定义了圆的中心位置，r 表示圆的半径。Hough变换还可以推广用来检测任意形状。但是，Hough变换存在一个问题，即检测简单曲线时需要大型参数搜索空间。这个问题可以通过附加信息来得到缓解，这些附加信息在空间域中可能有效。例如，通过平滑法来检测圆形曲线时，必须在三维参数空间中进行搜索。但是，当曲线斜率的方向已知时，搜索过程就只需要在一维空间中进行了。

（4）模式分类

从图像中提取出来的特征可以用一个多维特征空间来表示，而分类器可以用来识别目标（如最小距离分类器，该分类器被广泛用于统计模式识别过程中）。这种模式分类方法在那些目标为图像的某个区域指定一个标签（有很多标签中的一个）的应用中尤其有效。分类器可能会以有监督学习的模式出现（利用已知目标类型的原型），或者以无监督学习的模式出现；在无监督模式中，学习过程是自动进行的。选择合适的特征对于这种方法的成功至关重要。人们已经开发出了很多利用结构关系的模式分类方法。关于模式分类的完整讨论，读者可以参考Duda和Hart（1973）的相关书籍。

19.5 三维目标识别

我们观察和接触到的现实世界主要是由三维的固定目标构成的；当对一个目标进行初次观察之后，人们就会从各个不同的角度来收集该目标的信息。收集目标的详细信息并保存信息的过程称为“模型建立”。一旦一个人对很多目标都很熟悉之后，这些目标就可以从任意的角度来进行识别，而无需进行进一步的研究。人们也可以识别、定位和定性地描述目标在黑白图片中的方位。这种基本能力对于机器视觉系统非常重要，因为它涉及到了线框矩形区域中的单个参数空间变量，该矩形区域对应了一个固定的、单视角的世界。

1. 识别系统的构成

识别过程是指辨认已知的事物。为了达到识别的目的，在实际目标的建模过程中，可以采用各种不同的方案。为了确定如何进行识别，必须采用传感器数据与模型数据进行匹配的方法。简单的盲目搜索方法可以实现这个过程，即：将所

有可区别方向和位置的已知目标模型的所有可能组合转换成数字化的传感器格式；基于最小匹配误差规范进行匹配。很显然，这种方法是不切实际的。另一方面，由于目标模型中包含了比传感器数据更多的目标信息，因此我们无法将传感器数据转换成完整的模型数据，也无法匹配成模型数据格式。但是，这并不能阻止我们对部分模型数据进行匹配。因此，如果采用中间域来计算传感器数据和模型数据，就行得通了；这个中间域称为“符号场景描述域”。在该描述域中，匹配流程是以数量的形式进行的，该数量就是一个“特征”。

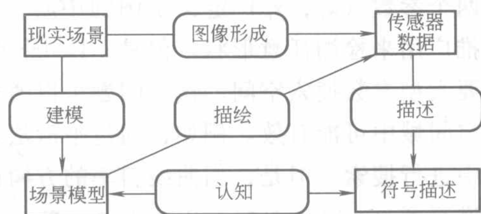


图 19-12 目标识别系统的构成（来源：Best，P. and Jain. R. C. 1985. Three dimensional object recognition. *ACM Computing Surveys* 17(1).）

图 19-12 描述了识别系统中各个组成部分之间的相互关系。图像信息处理过程可以产生完全遵循物理规律的亮度数据或距离数据。描述过程负责处理传感器数据并提取与应用无关的特征。这个过程完全是数据驱动的，而且只涉及到图像形成过程的相关知识。建模过程负责提供实际目标的目标模型，根据传感器数据来进行目标重构是自动建模的一种方法。认知过程或识别过程包含了模型和数据描述法之间的匹配实现机制，这个过程可能包含了数据驱动和模型驱动子处理过程；在这些子过程中，分段传感器数据区域会根据模型来寻找详细说明，而且假定的模型会从数据中寻找识别信息。描绘过程负责根据目标模型来生成综合的传感器数据，描绘过程还通过一个自主系统来提供一个重要的反馈连接，以便通过比较综合图像和传感器图像之间的区别来核对自身关于传感器数据的识别结果。

2. 目标表示法

人们开发出了各种各样的三维目标表示法，而选择合适的方法是由应用本身决定的。在计算机制图中，线框法和结构立体几何法是很受欢迎的，因为它们的数据结构非常适合对图像的描绘。在机器视觉系统中，其他应用广泛的目标表示法包括广义代码法和特征视角法。接下来，我们将主要描述一些常用的目标表示法。

（1）线框法

三维目标的线框表示法是由一个三维顶点序列和一个边缘顶点对序列构成的。尽管这种表示法很简单，但是当用来确定表面面积和目标大小时，它也是一种不明确的表示法。线框模型有时可以看做是几个不同的立体目标，或者相同目标的不同方位。图 19-13a 给出了一个简单的三维目标的线框表示法示例。

（2）结构立体几何法

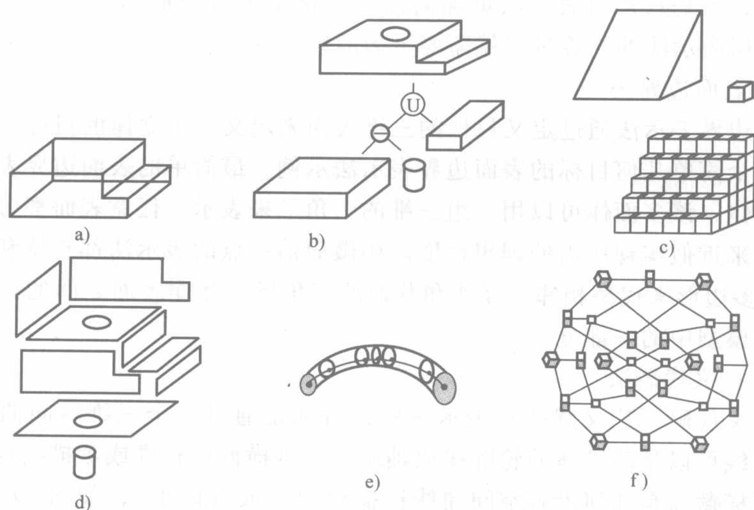


图 19-13 各种目标表示法

a）线框法 b）结构立体几何法 c）空间堆积法

d）表面边界法 e）广义代码法 f）视点图形法

目标的结构立体几何（Constructive Solid Geometry，CSG）表示法是按照一组三维立体单元（如块、圆柱、锥、球等等）和一组布尔算子（联合、交叉和差异）来进行描述的。图 19-13b 给出了一个简单几何目标的 CSG 表示法示例。存储数据结构为二进制树形结构；在该结构中，终端节点是指各个几何单元，而分支节点代表了布尔算子和位置信息。CSG 树形结构明确定义了目标表面的面积，而且可以用很少的数据来表示复杂的目标。但是，那些要求获得有效表面信息的边界评估机制的计算量就非常大。同样，一般有刻纹的表面也很难用 CSG 模型来表示。

(3) 空间堆积法

空间堆积法利用目标中不重叠的三维空间子区域来定义目标。这种方法明确定义了目标的体积。这类方法中常见的原始表示法包括三维像素法和八叉树表示法。三维像素就是离散三维空间中的小型体积单元，如图 19-13c 所示。这些体积单元通常是指固定大小的立方体。目标就是通过堆积的一系列三维像素来表示的。三维像素法对存储器要求非常高，但是使用规则却非常简单。八叉树表示法是一个空间堆积的分级表示法。目标的体积会被划分成不同大小的立方体，立方体的大小取决于与根节点的距离。树形结构中的每个分支节点代表了一个立方体，并指向其他 8 个节点，这 8 个节点中的每一个节点用来描述目标的体积，这个体积对应了分支节点立方体中的八分体。八叉树表示法可以提供三维像素法的

所有优势，但是八叉树表示法更加简洁；正是由于这个原因，与三维像素法相比，八叉树表示法的计算过程就需要更多的复杂算法。

（4）表面边界法

表面边界表示法通过定义目标的三维表面来定义一个立体的目标。图19-13d给出了一个简单几何目标的表面边界表示法示例。最简单的表面边界表示法是三角形多面体，该多面体可以用一组三维的三角形来表示。任意表面都可以通过多个三角形来近似实现所需的理想精度。稍微简洁一点的表示法都支持利用任意的n边平面多边形来代替相邻、互连和共面的三角形。互连表面之间的结构关系也可以作为模型中的一部分。

（5）广义代码法

在广义代码（广义圆柱）表示法中，目标是通过一个三维空间曲线来描述的，该曲线可以是圆锥体的轮廓线或轴线、二维横截面轮廓或者描绘规则，该规则定义了横截面是如何沿着空间曲线被描绘的（或者修正），如图19-13e所示。广义代码法非常适合用来描述很多实际的目标形状。但是，有些目标是很难用广义代码法来描述的（如人的面孔或汽车车身）。如果忽略这些局限性的话，广义代码表示法在机器视觉系统中也是很受欢迎的。

（6）构架表示法

构架表示法采用空间曲线构架来表示目标。一个构架可以看作是广义代码描述的抽象，该描述只由轮廓构成。构架几何学提供了各种实用的抽象信息。如果构架上的每个点都指定一个半径函数，那么这种表示法就可以用来描述常用目标。

（7）多重二维投影法

在某些应用中，我们可以很方便地保存描述三维目标的二维轮廓投影库。对于平台上具有少量稳定方位的三维目标识别过程来说，如果目标轮廓足够清晰的话，那么多重二维投影就是最理想的表示法。例如，轮廓可以用来从光洁的天空背景中识别任何方位的飞行器。但是，由于很多不同的三维目标形状具有相同轮廓投影，因此，该类型的表示法不是一种通用的技术。

（8）视点图形法

在视点图形表示法中，观察点的空间被划分成各个极大的区域，在这些区域中，每个区域中的每个观察点都会给出定性的目标视觉，称为“视点”。在每个区域中，目标的投影具有相同数量和相同类型的特征，以及相同的空间关系。但是，这些特征的定量属性（如边缘的长度）会随着视角的不同而发生变化。视点的变化（称为“视觉事件”）主要发生在区域之间的边界上。如果两个视点对应的区域在视角空间中是相邻的话，那么这两个视点可以认为是通过视觉事件相互连接的。一个视点图就是一个图形结构，该图形结构中的节点代表了视点和节

点之间的关联区域，弧线代表了视觉事件和相邻区域之间的边界。图19-13f给出了一个立方体的视点图形示例。

（9）特征视角法

与视点图形法非常类似的概念就是特征视角法。在特征视角法中，目标的所有无限个二维立体投影点都会被划分成有限个拓扑等价类型。等价类型中不同投影点之间是通过线性变换来相互关联的。等价类型的描述性构成单元称为该类的“特征视角”。在特征视角法中，通常认为目标是停留在一个支撑面上，因此，目标就只能出现在稳定的位置上。目标的特征视角是由摄像机配置上的某些约束条件决定的。正是因为使用了摄像机位置和方向信息，我们才能将特征视角表示法与视点图形法区分开。由于特征视角法指定了目标的三维结构，因此该方法也可以做为常用的目标表示法。

19.6 动态视觉

早期的机器视觉系统关注的主要是静态场景，但是现实场景是动态的，因此，设计能分析动态场景的机器视觉系统就越来越受关注了。对于执行重要目标性能操作和任务的机器视觉系统来说，应对移动变化目标和视角的能力就显得异常关键了。

动态场景分析系统输入的是一系列的图像帧。用来获取图像序列的摄像机自身也可能处于移动的状态。每一帧图像代表了一幅瞬间的场景图，导致场景图发生变化的原因可能包括摄像机的移动、目标的移动、照射点的变化或者目标结构、大小、形状的变化。一般来说，场景图的变化通常认为是由摄像机或目标的移动导致的，我们一般认为目标是固定的或者准固定的。其他变化是不允许的。

一个场景中通常包含多个目标。某个指定时间点的场景图像描述了场景的部分投影，这部分投影与摄像机的位置有关。以下4种情况可以代表动态摄像机/目标所有可能的情形：

1）固定摄像机/固定目标（Stationary Camera/Stationary Objects，SCSO）；

2）固定摄像机/移动目标（Stationary Camera/Moving Objects，SCMO）；

3）移动摄像机/固定目标（Moving Camera/Stationary Objects，MCSO）；

4）移动摄像机/移动目标（Moving Camera/Moving Objects，MCMO）。

第一种是简单的静态场景分析情形。在很多应用中，处理单个图像来获取所需的信息是切实可行的。但是，更多的应用需要从动态环境中提取所需的信息。

很明显，一系列的图像帧可以提供更多的信息来帮助我们分析场景，但同时也给数据处理增加了更大的工作量。对每一个序列帧进行静态场景分析时需要庞

大的计算量，而且还会产生很多静态场景分析问题。幸运的是，关于动态场景分析的研究表明信息复原在动态场景中比静态场景中更加容易。在有些情况下，总体计算量可能更小，而且性能更好。

SCMO 场景是动态场景分析中关注最多的情形。SCMO 场景分析的目的通常是为了检测移动目标的运动并提取外部特征，从而来识别它们并计算它们的移动特性。MCMO 是最常见的情形，它可能代表了动态场景分析中最复杂的情形。人们开发出了很多技术来分析 MCMO 的情形，这些技术假设摄像机是固定的，因此不适用于移动的摄像机情形。类似的，为移动摄像机开发的技术通常假设场景是静态的，因此无法适用于动态目标。SCMO 和 MCSO 已经应用到了很多系统中，而且研究人员还对各种假设条件和各种背景下的应用进行了研究。

1. 变化检测

场景中任何可察觉到的运动都会在场景的一系列图像帧中得以体现。如果这些变化被检测出来，那么就可以对其运动特征进行分析了。如果一个目标的运动限制在一个平面上，该平面与图像面平行，那么我们就可以得到一个关于该目标运动构成的良好定量评估。对于三维运动来说，只存在定性的评估。通过分析帧与帧之间的区别，就可以对序列进行整体分析。变化可以从不同的等级来检测，这些等级包括：像素、边缘和区域。

（1）微分图

检测两个帧之间变化最简单的方法就是直接比较帧对应的像素，以确定它们之间是否相同。最简单的形式就是，我们可以通过计算下面的公式，来得到第 j 帧与第 k 帧之间的二进制微分图 $DP_{jk}(x,y)$：

$$DP_{jk}(x,y)=\begin{cases}1; & \text{如果}\ |F(x,y,j)-F(x,y,k)|>\tau \\ 0; & \text{否则}\end{cases} \tag{19-18}$$

式中，τ 是指门限值；$F(x,y,j)$ 和 $F(x,y,k)$ 分别是指第 j 帧和第 k 帧的图像阵列。

在微分图中，值为 1 的像素会被认为是目标运动的结果。但是，由于干扰，这种对真实场景的简单测试通常会产生令人不满意的结果。一个简单的尺寸过滤器就可以用来过滤掉那些没有形成最小互连串的像素。然后，只有那些微分图中值为 1 的像素才是由目标的运动产生的，这些值为 1 的像素属于 4 个互连单元（或者 8 个互连单元），这些互连单元比一组像素要大。这种过滤器在减小干扰方面非常有效，但是它也会过滤掉一些理想的信号，如来自小型目标或缓慢移动目标的信号。

（2）似然比

为了使目标运动变化的检测更加准确，必须更加充分地考虑两个帧中相同位置的像素区域或像素群，并严格比较它们之间的亮度特性。一种方法就是利用似

然比来比较各个帧。因此，我们可以利用 λ 代替 $|F(x,y,j)-F(x,y,k)|$ 来计算微分图，其中，

$$\lambda=\frac{\left[\dfrac{\sigma_1^2+\sigma_2^2}{2}+\left(\dfrac{\mu_1-\mu_2}{2}\right)^2\right]^2}{\sigma_1^2\sigma_2^2} \tag{19-19}$$

式中，μ 和 σ 分别代表了采样区域的平均灰度值和帧方差的平方根。

似然比只能应用于区域中，不能应用于单个像素。这个限制条件带来了一点小问题，不过这个小问题可以通过假设帧的对应区域来解决。似然比测试与尺寸过滤器结合起来可以很好地消除很多真实场景中的干扰。通过假设图像中每个像素中心的重叠区域，似然比测试就可以应用于每个帧中的每个点了，或者似然比测试可以通过使用不重叠的区域来实现对各个点的二次采样，这些不重叠的区域称为“超级像素”。

(3) 累积差分图

小型目标或缓慢移动目标的检测遗漏问题可以通过分析整个帧序列（而不是两个帧之间）的变化来解决。累积差分图（Accumulative Difference Pictures, ADP）就是用来检测小型目标或缓慢移动目标的。累积差分图是通过比较图像序列中的每个帧与共同参考帧之间的区别来形成的。如果对应区域的似然比超出了门限值，那么累积差分图的输入值就会增加1。因此，一个包含超过 k 个帧的累积差分图可以由下式表示：

$$ADP_0(x,y)=0$$

$$ADP_k(x,y)=ADP_{k-1}(x,y)+DP_{1k}(x,y) \tag{19-20}$$

一个帧序列中的第一个帧通常作为参考帧。

(4) 时变边缘检测

由于静态场景中的边缘检测非常重要，因此我们期望时变边缘检测在动态场景的分析中也同样重要。我们可以利用逻辑 AND 运算并通过合并空间和时间斜率来检测移动边缘；其中，逻辑 AND 运算可以通过乘法器来实现。因此，图像帧中某个点的时变边缘性可以由式（19-21）表示：

$$E_t(x,y,t)=\frac{\mathrm{d}F(x,y,t)}{\mathrm{d}S}\,\frac{\mathrm{d}F(x,y,t)}{\mathrm{d}t} \tag{19-21}$$

式中，$\mathrm{d}F/\mathrm{d}S$ 和 $\mathrm{d}F/\mathrm{d}t$ 分别是指点（x,y,t）处像素亮度的空间斜率和时间斜率大小。

各种常规的边缘检测器可以用来计算空间斜率，而简单微分可以用来计算时间斜率。在大多数情况下，这种边缘检测器非常有效的。通过对上面的乘积进行门限处理，可以克服小型目标或缓慢移动目标的检测遗漏问题（而不是首先进行一阶微分处理，然后使用边缘检测器；也不是首先检测边缘，然后计算它们的

时间斜率)。

2. 光学流

光学流是指图像中所有点的速度分布，该分布与观察者有关。光学流承载了动态场景分析中的各种有效信息。光学流是由图像中每个点的速度矢量决定的。人们设计了很多方法用来计算基于两个或多个帧序列的光学流，这些方法可以分为两大类：基于特征的方法和基于斜率的方法。如果使用的是固定摄像机，那么图像帧中的大多数点都将是零速度。光学流假设场景中的微小部分是移动的，而通常这也是真实的情况。因此，大多数光学流的应用都涉及到了移动的摄像机。

(1) 基于特征的方法

基于特征的光学流计算方法首先从连续的图像帧中选择一些特征，并在帧之间对这些特征进行匹配，然后计算它们之间的区别。如前所述，"对应问题"可以通过张驰技术来解决。但是，选择特征和建立对应关系的问题就不容易解决了。另外，这种方法只能为图像中很少的点产生速度矢量。

(2) 基于斜率的方法

基于斜率的方法定义了图像亮度空间斜率和时间斜率之间的关系。这种关系可以通过图像点的速度来分割图像。空间斜率和时间斜率以及速度分量之间的关系如下：

$$F_x u + F_y v + F_t = 0 \tag{19-22}$$

式中，$u = \mathrm{d}x/\mathrm{d}t$，$v = \mathrm{d}y/\mathrm{d}t$。

在该方程式中，F_x、F_y、F_t分别为x方向和y方向的空间斜率以及时间斜率，它们可以从图像中直接计算得到。这样，图像中的每一个点都有两个未知量u和v，但只有一个等式，因此，就无法直接确定光学流。

但是，我们可以假设速度域在整个图像中是平稳变化的。基于这种假设，就可以利用两个或多个帧，采用一种迭代的方法来计算光学流。下面给出了用来计算光学流的迭代公式：

$$u = u_{平均} - F_x \frac{P}{D} \quad v = v_{平均} - F_y \frac{P}{D} \tag{19-23}$$

式中，

$$P = F_x u_{平均} + F_y v_{平均} + F_t \tag{19-24}$$

$$D = \lambda^2 + F_x^2 + F_y^2 \tag{19-25}$$

式中，λ是指常数乘法器。

如果只有两个帧，那么上面的迭代计算过程就会在相同帧反复进行多次。如果包含多个帧，那么每一次迭代计算过程都会使用一个新的帧。

3. 利用移动摄像机进行图像分割

如果摄像机是移动的，那么图像中的每个点就会具有与摄像机相关的非零速

度（除非目标点的移动速度与摄像机完全相同）。与摄像机相关的速度取决于点本身的速度，也取决于点与摄像机之间的距离。基于微分的方法可以扩展到移动摄像机场景的分割过程中。如果我们的目标是为了提取移动目标的图像，那么就需要额外的信息来确定一个点的运动是否只与其距离有关，还是与其距离和运动同时有关。基于斜率的方法同样也需要这样的额外信息。

如果摄像机的运动方向是已知的，那么场景中固定分量的伸展中心（Focus Of Expansion，FOE）就可以很容易计算出来。图像平面中的 FOE 坐标（x_f，y_f）为

$$x_f = \frac{dx}{dz} \quad y_f = \frac{dy}{dz} \tag{19-26}$$

图像中所有固定点的速度矢量都投影在图像平面上，并相交于 FOE 处。根据 FOE 的变换可以用来简化分割过程。一个图像的自我运动极性（Ego-Motion Polar，EMP）变换通过以式（19-27）将帧 $F(x,y,t)$ 变换成 $E(r,\theta,t)$：

$$E(r,\theta,t) = F(x,y,t) \tag{19-27}$$

式中，

$$r = \sqrt{(x-x_f)^2 + (y-y_f)^2} \tag{19-28}$$

$$\theta = \tan^{-1}\left[\frac{(y-y_f)}{(x-x_f)}\right] \tag{19-29}$$

在 EMP 空间中，固定点是沿图像序列帧之间的 θ 轴分布的，而移动目标上的点是沿 r 轴和 θ 轴分布的。因此，EMP 的空间分布可以用来将场景分割成固定和移动部分。另外，当对 FOE 进行复杂对数映射（Complex Logarithmic Mapping，CLM）时，就可以得到 EMP 空间中的“兴趣”点。

19.7 应用

机器视觉系统已经广泛应用于各个领域中，例如从医学到机器人技术，从自动检查到自主导航，从文档分析到多媒体系统等。新的机器视觉系统也在不断涌现，而且越来越成为我们生活中不可缺少的一部分。在本节中，我们将主要对机器视觉的各种应用进行简要介绍。

1. 光学字符识别和文档图像分析

文档图像分析的目的是为了识别图像中的文字和图形，并提取出我们需要的信息。文档处理过程可以分为两种类型：文字处理和图形处理。文字处理主要负责处理文档图像中的文字成分，其中一部分任务包括通过光学字符识别（Optical

Character Recognition，OCR）技术来识别文字、扭曲检测和纠正（文档扫描时产生的倾斜）、发现纵列、图形、文字线以及文字。图形处理主要负责处理文档图像中的非文字成分，如构成线路图的线条和符号，并确定文字段落之间以及内部标识语之间的界限。

文档分析是目前非常活跃的研究领域。研究人员已经开发出了各种文档分析系统，如表格图的自动工程制图说明和识别系统、邮政信封上的邮政区号识别以及乐谱说明系统等等。但是，文档分析真正成功的范例还是 OCR 系统。OCR 是机器视觉系统中的一个技术领域，该领域中的各种科学研究已经诞生出了很多低成本的市场产品。目前，很多 OCR 系统的识别准确性已经超过了 90%。

在本章最后的备注中列出了很多关于文档分析的期刊和会议学报。另外，"国际文档分析与识别会议（International Conference on Document Analysis and Recognition，ICDAR）"和"国际图形识别研习会（International Workshop on Graphics Recognition，IWGR）"每年会联合举办两次，这两个会议对文档分析领域的发展起着重要的推动作用。

2. 医学图像分析

医学图像分析主要负责处理诸如 X 射线、计算机断层造影（Computerized Tomography，CT）扫描和核磁共振成像（Magnetic Resonance Imaging，MRI）等图像。早期的医学图像分析和图像处理是相互重叠的，其主要任务是辅助医师观察医学图像，其中不包含任何自动说明或高水平的系统推论。不过，目前很多最新的成果已经被引入到了医学成像技术当中，医学成像是最符合机器视觉定义的技术。例如，基于已知模型（图像）和患病样本的特征来搜索患病器官或组织图像的系统、基于 CT 扫描和 MRI 来生成器官三维模型的系统。目前比较活跃的医学图像研究领域包括在外科手术和手术方案中（尤其是神经外科）使用三维图像、虚拟手术以及根据图像序列生成时变三维模型（如起搏的心脏）等。这些研究领域涉及到了机器视觉系统中的很多方面，同时还结合了很多计算机制图技术。

除了本章结束后列出的参考文献外，下面还给出了关于当前医学成像领域研究的一些宝贵资料："IEEE 医学成像学报（*IEEE Transactions on Medical Imaging*）"、"IEEE 生物医学工程学报（*IEEE Transactions on Biomedical Engineering*）"、"IEEE 医学图像处理学报（*IEEE Transactions on Medical Image Processing*）"和"IEEE 医学和生物工程杂志（*IEEE Engineering in Medicine and Biology Magazine*）"。同样，还有一些为医学成像研究的发展起重要推动作用的会议："SPIE 医学成像（SPIE Medical Imaging）会议"、"IEEE 医学和生物工程（IEEE Engineering in Medicine and Biology）会议"、"虚拟现实医学（Medicine Meets Virtual Reality）会议"、"IEEE 可视化精选座谈会（select sessions/symposium of

IEEE Visualization）”和“生物医学计算的 SPIE 可视化（SPIE Visualization in Biomedical Computing）会议”。

3. 摄像测量法和航空图像分析

摄像测量法主要负责通过图像制定可靠的测量方法。在早期的摄像测量法中，图像实际上是从气球上拍摄打印出来的照片。不过，目前的遥感处理过程利用了多谱成像技术，该过程使用了很多其他电磁频谱的能量，如紫外线和红外线。通常，很多图像都是直接从围绕地球运动的卫星上发送过来的，如在 20 世纪 70 年代初发射的美国 Landsat 卫星。摄像测量法和航空图像分析的应用包括大气探测、能量守恒定律的热成像分析、对自然资源、农作物状况、土地覆盖和土地使用情况的监控、气象预报、污染调查、城市规划、军事侦察以及其他在地质学、水文地理学和海洋学等方面的应用。

很多组织机构为这些遥感技术的研究做出了重要贡献。下面给出了关于摄像测量法和航空图像分析方面的重要参考资料："美国摄像测量协会（*The American Society of Photogrammetry*）”、“国际遥感学会（*The International Remote Sensing Institute*，*ISRI*）”和“美国测量和绘图大会（*The American Congress on surveying and Mapping*）”。另外，还有很多关于这方面主题的会议：“SPIE 遥感会议（*SPIE Remote Sensing*）”、“国际地球科学和遥感座谈会（*International Geosciences and Remote Sensing Symposium*）”、“IEEE 地球科学和遥感会议学报（*IEEE Transactions on Geosciences and Remote Sensing*）”以及由美国地质协会主办的遥感技术座谈会。

4. 自动检查和机器人技术

与前面提到的机器视觉系统应用不同，自动检查和机器人技术主要完成一些实时任务，这些实时任务在很大程度上增加了系统的复杂度。这些系统的目标就是实现传感器指导控制。典型的工业应用包括机械零件、焊接面（焊接缝）、硅片、农产品甚至糖果的自动检查。工业视觉系统设计师经常面临的挑战包括：确定摄像机和光照的最佳配置、确定最合适的彩色照射空间描述方式、对各种表面反射现象进行建模、动态传感器反馈、实时的操作控制、实时的操作系统接口以及神经网络。

结构化光源技术已经广泛应用于工业机器视觉系统中，其中场景照射的控制非常容易。在典型的应用中，传送带上的目标通过一个光照面时，会在图像中产生一个光条纹扭曲；然后，我们就可以计算光束面上的目标轮廓了。当目标沿着传送带移动时，上面的过程以一定的间隔反复进行，从而覆盖到目标的整个形态。之后的操作就根据系统的目的进行了。

下面给出了关于工业机器视觉系统和机器人技术应用方面的主要参考资料：“国际机器人研究杂志（*International Journal of Robotics Research*）”、“IEEE 机器人技术和自动化学报（*IEEE Transactions on Robotics and Automation*）”、“IEEE 国

际机器人技术和自动化会议（*IEEE's International Conference on Robotics and Automation*）”和“*IEEE* 系统、人和控制论学报（*IEEE Transactions on Systems, Man, and Cybernetics*）”。

5. 自主导航

与机器人技术接近的就是自主导航领域。为了使机器人或其他移动机械能够在特殊的环境中自动导航，人们作了大量的努力。其中涉及到的技术包括有源视觉（传感器控制）、神经网络、高速立体视觉、三维视觉（距离成像）、高水平导航方案推论以及精确的远程机械控制信号压缩与传输。

6. 视觉信息管理系统

最新的机器视觉技术研究领域可能就是视觉信息管理系统（Visual Information Management System，VIMS）。随着低成本计算机和多媒体技术的应用，数字影像成为了我们日常生活中越来越不可缺少的一部分。VIMS 的研究可以为我们提供处理这些数字信息的方法。VIMS 应用包括：互动电视、视频电话会议、数字图书馆、视频点播以及大规模视频数据库。图像处理和机器视觉技术在以下这些应用中也同样扮演了重要角色，如视频压缩方案的设计（允许多种技术直接处理数据流，并为多维数据开发更有效的寻址方法）、场景缺口的自动检测（以便自动寻址更大的视频数据存储空间并开发出根据图像内容来查询图像数据库的方法），该自动检测技术与标准的结构化查询语言（Structured Query Language，SQL）技术原理恰好相反。

名词解释

对应问题：一个图像中的点与另一个图像中的点之间的匹配问题。

矩形图：图像中灰阶出现的频繁度示意图。

量化：通过有限个离散灰阶来描述图像亮度连续变化范围的过程。

采样：将连续的图像亮度函数表示成离散的二维阵列的过程。

分割：将目标从背景中区分出来的过程。

多边形法：通过一组互连的直线段来表示轮廓线的方法。对于闭合曲线来说，这些直线段可以形成一个多边形。

投影：将多维空间变换成少维空间的过程（例如，将三维空间场景变换成二维图像）。

门限处理：通过选择一个区间（通常在像素亮度上选择）来将目标从背景中区分出来的方法，亮度值在区间之内的点将其像素值设置为“1”，亮度值在区间之外的点将其像素值设置为“0”。

参考文献

[1] Besl, P. J. 1988. Active, Optical range imaging sensors. *Machine Vision and Applications* 1 (2): 127-152.

[2] Besl, P. and Jain, R. C. 1985. Three dimensional object recognition. *ACM Computing Surveys* 17 (1).

[3] Duda, R. O. and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*. Wiley, New York.

[4] Foley, J. D., van Dam, A., Feiner, S, K., and Hughes, J. F. 1990. *Computer Graphics, Principles and Practice*. Addison-Wesley, Reading, MA.

[5] Gonzalez, R. C. and Woods, R. E. 1992. *Digital Image Processing*. Addison-Wesley, Reading, MA.

[6] Javis, R. A. 1983. A perspective on range finding techniques for computer vision. *TEEE Trans. on Pattern Analysis and Machine Intelligence* 5 (2): 122-139.

[7] Kasturi, R. and Jain, R. C. 1991. *Computer Vision: Principles*. IEEE Computer Society Press.

[8] Kosko, B. 1992. *Neural Networks and Fuzzy Systems*. Prentice-Hall, Englewood Cliffs, NJ.

[9] Jain, R., Kasturi, R. and Schunck, B. G. 1995. *Machine Vision*. McGraw-Hill, New York.

[10] Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco, CA.

[11] Tanimoto, S. L. 1995. *The Elements of Artificial Intelligence Using Common Lisp*. Computer Science Press.

[12] Winston, P. H. 1992. *Artificial Intelligence*. Addison-Wesley, Reading, MA.

备注

本章中很多论点都来自于以下资料：

[1] Jain, R., Kasturi, R. and Schunck, B. G. 1995. *Machine Vision*. McGraw-Hill, New York.

[2] Kasturi, R. and Jain, R. C. 1991. *Computer Vision: Principles*. IEEE Computer Society Press.

另外，向读者推荐以下参考书籍：

[1] Rosenfeld, A. and Kak, A. C. 1982. *Digital Picture Processing*. Academic Press, Englewood Cliffs, NJ.

[2] Jain, A. K. 1989. *Fundamentals of Digital Image Processing*. Prentice-Hall, New York.

[3] Haralick, R. M. and Shapiro, L. G. 1992-1993. *Computer and Robot Vision*. Addison-Wesley, Reading, MA.

[4] Horn, B. 1986. *Robot Vision*. McGraw-Hill.

[5] Schalkoff, R. J. 1989. *Digital Image Processing and Computer Vision*. Wiley, New York.

关于机器视觉技术的信息，读者可以参考以下技术期刊：

[1] “人工智能（*AI*, *Artificial Intelligence*）”

[2] “机器视觉、图形和图像处理（*CVGIP*, *Computer Vision*, *Graphics*, *and Image Processing*）”

[3] “IEEE 图像处理学报（*IEEE Transactions on Image Processing*）”

[4] “IEEE 模式分析和机器智能学报（*IEEE Transactions on Pattern Analysis and Machine Intelligence*）”

[5] “图像和视觉计算（*Image and Vision Computing*）”

[6] “国际机器视觉杂志（*International Journal of Computer Vision*）”

[7] “国际模式识别和人工智能杂志（*International Journal on Pattern Recognition and Artificial Intelligence*）”

[8] “机器和应用（*MVA*, *Machine and Applications*）”

[9] “模式识别（*PR*, *Pattern Recognition*）”

[10] “模式识别信函（*PRL*，*Pattern Recognition Letters*）”

下面关于机器视觉技术的会议学报也是很好的参考资源：

[1] 由国际光学工程协会（International Society for Optical Engineering）举办的一系列会议。
[2] 由电气和电子工程师学会（IEEE，Institute of Electrical and Electronic Engineering）举办的一系列研讨会。
[3] 由国际模式识别联盟（IAPR，International Association for Pattern Recognition）举办的一系列研讨会。
[4] “机器视觉欧洲会议（ECCV，European Conference on Computer Vision）”。
[5] “IEEE 机器视觉和模式识别会议（CVPR，Conference on Computer Vision and Pattern Recognition）”。
[6] “国际机器视觉会议（ICCV，International Conference on Computer Vision）”。
[7] “国际模式识别会议（ICPR，International Conference on Pattern Recognition）”。

第20章 语音增强技术概述[一]

Yariv Ephraim　Hanoch Lev-Ari　William J. J. Roberts

20.1 引言

语音增强技术的目的是为了提高语音通信系统在吵杂环境中的性能。语音增强技术可以应用于移动无线通信系统、语音识别系统、低质量的录音系统，也可以用来提高助听器的性能。干扰源可能是宽带噪声，并以白噪声或有色噪声、周期信号如交流声、室内混响的形式出现，也可能表现为衰减噪声的形式。前面两种是代表了相加性噪声源，而后两种分别代表了卷积噪声源和乘法噪声源。语音信号也可能同时受到多个噪声源的干扰。

衡量语音增强系统主要有两种感性标准。增强信号的“质量”是指它的清晰度、失真类型和信号中残留的噪声强度。质量只是一种主观的衡量方法，该方法表示了收听者对增强信号的适应程度。第二种标准是指衡量增强信号的“可理解性”，这是一种客观的衡量方法，它给出了收听者可以正确辨别文字的百分比，这种方法中文字不需要很有意义。这两种性能衡量方法之间是不相关的。一个信号可能具有较好的质量，但可能具有较差的可理解性；或者可能具有较好的可理解性，但可能具有较差的质量。大多数语音增强系统都选择提高信号的质量，而牺牲信号的可理解性。收听者通过仔细收听信号，通常可以在噪声中提取出比增强信号更多的信息。这个观点显然来自于信息论中的数据处理定律。但是，收听者在经过长时间的收听对话之后会产生疲劳，因此就会降低对噪声信号的可理解性。在这种情况下，增强信号的可理解性就可能比噪声信号的可理解性高了。这样，收听者不需费多大的劲就可以区别出增强信号了，这些增强信号对应了噪声信号中的高信噪比部分。

质量和可理解性都很难衡量，因为它们要求收听对话是关于生活主题的。这样，研究人员通常会采取准规范的收听测试方式来评定一个增强信号的质量，同时利用自动语音识别系统测试方式来评定该增强信号的可理解性。质量和可理解

[一] 本章中的部分内容来自 Hanoch Lev-Ari 和 Yariv Ephraim 的 “Extension of the Signal Subspace Speech Enhancement Approach to Colored Noise”，这篇文章刊登在 IEEE Sig. Proc. Let.，vol. 10，pp. 104-106，April 2003©2003 IEEE。

性同样也很难量化，也很难以完美的形式（可以用来进行数学优化）进行表达。因此，语音增强系统的设计通常是基于数学衡量标准的，该衡量标准被认为与语音信号的质量和可理解性有关。一个典型的例子就是通过最小化原始信号和评估信号之间的均方误差（Mean Square Error，MSE）来对清晰信号进行评估[5]。这种标准被认为比最小化原始信号和评估信号波形之间的 MSE 在感性上更加有意义[13]。

高效语音增强系统设计中的另一个难点就是缺乏明确的语音信号和噪声过程统计模型。另外，语音信号（也可能包含噪声过程）不是一个严格意义上的稳定过程。语音信号的常用参数模型，如信号短期建模的自回归过程和信号长期建模的隐蔽马尔可夫过程（Hidden Markov Process，HMP），都不能为语音增强系统提供合适的模型。Lim 和 Oppenheim 开发出了不同的最大期望值（Expectation-Maximization，EM）算法，并应用到了语音增强技术中[12]；该算法用来计算噪声信号中自回归参数的极大似然（Maximum Likelihood，ML）估计值。许多评估方案在很多年前就已经被开发出来了，例如 Ephraim[6]；这些方案都基于清晰语音信号的隐蔽马尔可夫建模和噪声过程的隐蔽马尔可夫建模。在每一种情况中，语音信号和噪声过程的 HMP 分别是根据两个过程的训练序列来设计的。实践发现，自回归隐蔽马尔可夫模型在编码和清晰语音信号识别方面非常有效，但同时在语音增强应用的精炼中不是很有效。

在本章中，我们将重温那些主要用于相加性宽带噪声源的常用语音增强方法。尽管其中的有些方法是用来降低回响噪声的，但我们相信消除回响噪声问题需要一个完全不同的方法，该方法不属于本章介绍的范围。我们将主要关注频谱消减法[13]及其部分派生方法，如信号子空间法[7][11]，以及短期频谱估计法[16,4,5]。本章选择这几种方法作为主要介绍对象，是由于有些频谱消减法的派生方法仍然是目前最有效的方法。这些方法实现起来相对简单，而且通常比那些依靠参数统计模型和训练流程的方法更加有效。

20.2 信号子空间法

在本节中，我们将介绍信号子空间法的原理及其与维纳（Wiener）过滤法和频谱消减法之间的关系。我们的介绍从参考文献［7,11］中的观点开始。信号子空间法假设信号和噪声之间是不相关的，而且它们的二阶统计特性是有效的。该方法没有对信号和噪声过程的分布做出假设。

我们用欧几里德空间 $\Re^k$ 中的 k 维随机矢量 $\boldsymbol{Y}$ 和 $\boldsymbol{W}$ 来分别代表清晰信号和噪声。假设在一个定义恰当的概率空间中，每个随机变量的期望值都为零。另外，假设 $\boldsymbol{Z}=\boldsymbol{Y}+\boldsymbol{W}$ 用来表示噪声矢量；$\boldsymbol{R}_y$ 和 $\boldsymbol{R}_w$ 分别用来表示清晰信号和噪声过程

的协方差矩阵，其中，$\boldsymbol{R}_w$为正定矩阵。假设$\boldsymbol{H}$表示线性空间$\Re^{k\times k}$中的一个$k\times k$的实矩阵；$\hat{\boldsymbol{Y}}=\boldsymbol{HZ}$表示给定$\boldsymbol{Z}$时$\boldsymbol{Y}$的线性估计式。估计过程中的残余信号为：

$$\boldsymbol{Y}-\hat{\boldsymbol{Y}}=(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}-\boldsymbol{HW} \tag{20-1}$$

式中，$\boldsymbol{I}$通常表示单位矩阵。

为了简化表示法，我们最好不要明确给出单位矩阵的维数，这些单位矩阵的维数应该根据上下文关系来确定。在式（20-1）中，$\boldsymbol{D}=(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}$是指信号失真度；$\boldsymbol{N}=\boldsymbol{HW}$是指线性估计过程的残余噪声。我们用$(.)'$来表示实矩阵的转置矩阵或者复矩阵的共轭转置矩阵，而且下式

$$\bar{\in}_d^2=\frac{1}{k}\mathrm{tr}E\{\boldsymbol{DD}'\}=\frac{1}{k}\mathrm{tr}\{(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{R}_y(\boldsymbol{I}-\boldsymbol{H})'\} \tag{20-2}$$

表示平均信号失真功率，其中 tr {.} 表示一个矩阵中主对角线元素的和。类似的，我们用下式

$$\bar{\in}_n^2=\frac{1}{k}\mathrm{tr}E\{\boldsymbol{NN}'\}=\frac{1}{k}\mathrm{tr}\{\boldsymbol{HR}_w\boldsymbol{H}'\} \tag{20-3}$$

表示平均残余噪声功率。

在设置残余噪声功率$\bar{\in}_n^2$门限值的基础上，矩阵$\boldsymbol{H}$通过最小化信号失真功率$\bar{\in}_d^2$来进行估计。在给定α的前提下，其计算公式如下：

$$\min_H \bar{\in}_d^2\text{条件为}\bar{\in}_n^2\leqslant\alpha \tag{20-4}$$

我们假设用$\mu\geqslant 0$表示不等式约束的拉格朗日乘子（Lagrange Multiplier），那么假定$\boldsymbol{H}=\boldsymbol{H}_1$，最佳矩阵就是：

$$\boldsymbol{H}_1=\boldsymbol{R}_y(\boldsymbol{R}_y+\mu\boldsymbol{R}_w)^{-1} \tag{20-5}$$

矩阵$\boldsymbol{H}_1$可以通过以下过程实现。假设$\boldsymbol{R}_w^{1/2}$表示$\boldsymbol{R}_w$的对称正定均方根，$\boldsymbol{R}_w^{-1/2}=(\boldsymbol{R}_w^{1/2})^{-1}$；假设$\boldsymbol{U}$表示对称矩阵$\boldsymbol{R}_w^{-1/2}\boldsymbol{R}_y\boldsymbol{R}_w^{-1/2}$特征矢量的正交矩阵；$\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_k)$表示$\boldsymbol{R}_w^{-1/2}\boldsymbol{R}_y\boldsymbol{R}_w^{-1/2}$非负特征值的对角矩阵。因此，

$$\boldsymbol{H}_1=\boldsymbol{R}_w^{1/2}\boldsymbol{U}\Lambda(\Lambda+\mu\boldsymbol{I})^{-1}\boldsymbol{U}'\boldsymbol{R}_w^{-1/2} \tag{20-6}$$

当式（20-6）中的$\boldsymbol{H}_1$应用到Z域时，首先通过将$\boldsymbol{R}_w^{-1/2}$应用到Z域来对输入噪声进行白化处理。然后，就可以使用白化清晰信号的协方差矩阵对应的正交变换$\boldsymbol{U}'$了，而且变换信号可以通过对角维纳（Wiener）型增益矩阵来进行修正。

在式（20-6）中，只包含噪声的白化噪声信号的构成元素为零，这些构成元素的索引为$\{j:\lambda_j=0\}$。当噪声为白色噪声时，$\boldsymbol{R}_w=\sigma_w^2\boldsymbol{I}$，$\boldsymbol{U}$和$\Lambda$分别为特征矢量的矩阵和$\boldsymbol{R}_y/\sigma_w^2$的特征值。信号中零元素$\{j:\lambda_j=0\}$的存在意味着信号处于欧几里德（Euclidean）空间$\Re^k$的字空间中。同时，噪声的特征值全部等于σ_w^2，因此，噪声占据了整个$\Re^k$空间。这样，信号子空间法首先消除信号子空间

中的噪声成分，然后根据式（20-4）的标准来修正信号空间中的信号成分。

如果信号和噪声是“宽带敏感”的稳定信号，那么矩阵 $\boldsymbol{R}_y$ 和矩阵 $\boldsymbol{R}_w$ 就是 Toeplitz 矩阵，而且分别具有相关的功率谱密度 $f_y(\theta)$ 和 $f_w(\theta)$，其中，角频率 θ 的范围为［0，2π）；如果信号和噪声是渐近弱稳定信号，那么矩阵 $\boldsymbol{R}_y$ 和矩阵 $\boldsymbol{R}_w$ 分别表示了渐近 Toeplitz 矩阵，而且分别具有相关的功率谱密度 $f_y(\theta)$ 和 $f_w(\theta)$[10]。由于后面的情况相对来说更加常见，因此接下来我们将着重介绍渐近弱稳定信号和噪声，这样式（20-5）中的过滤矩阵 $\boldsymbol{H}_1$ 就成了渐近 Toeplitz 矩阵，其相关功率谱密度为

$$h_1(\theta)=\frac{f_y(\theta)}{f_y(\theta)+\mu f_w(\theta)} \tag{20-7}$$

式（20-7）就是清晰信号的无关联维纳滤波器，其中噪声强度可以通过式（20-4）中的约束因子 α 来进行调整。该滤波器通常通过两个功率谱密度的估计来实现。假设 $f_y(\hat{\theta})$ 和 $f_w(\hat{\theta})$ 分别表示了 $f_y(\theta)$ 和 $f_w(\theta)$ 的估计值，这些估计值可以从周期图或者平滑的周期图中得到。在这种情况下，滤波器公式为

$$h_1(\hat{\theta})=\frac{f_y(\hat{\theta})}{f_y(\hat{\theta})+\mu f_w(\hat{\theta})} \tag{20-8}$$

式中，$f_y(\hat{\theta})$ 的计算方式为

$$f_y(\hat{\theta})=\begin{cases} f_z(\hat{\theta})-f_w(\hat{\theta}) & \text{非负} \\ \varepsilon & \text{其他} \end{cases} \tag{20-9}$$

然后就可以得到清晰信号的频谱消减估计算法。常量 $\varepsilon \geqslant 0$ 通常称为“频谱基数”。通常，该估计算法中的 $\mu \geqslant 2$。

增强型过滤矩阵 $\boldsymbol{H}$ 也可以通过利用残余噪声的约束条件来进行设计。这种方法可以对残余噪声的频谱进行造型，以尽量减小感性效应。假设一组 k 维的实矢量或复矢量 $\{\boldsymbol{v}_i,\ i=1,\ \cdots,\ m\}$，$m \leqslant k$ 和一组非负常量 $\{\alpha_i,\ i=1,\ \cdots,\ m\}$。矢量组 $\{\boldsymbol{v}_i\}$ 用来将残余噪声变换成频域信号，常量组 $\{\alpha_i\}$ 作为这些频谱分量方差的最大值。矩阵 $\boldsymbol{H}$ 的计算公式为

$$\min_{\boldsymbol{H}} \bar{\in}_d^2 \text{条件为} \quad E\{|\boldsymbol{v}_i{}'N|^2\} \leqslant \alpha_i, \quad i=1,\cdots,m \tag{20-10}$$

当噪声为白色噪声时，矢量组 $\{\boldsymbol{v}_i\}$ 可以是 $\boldsymbol{R}_y$ 的特征值矢量组，而且沿着这些坐标矢量方向的残余噪声的方差就会受到限制。换句话说，矢量组 $\{\boldsymbol{v}_i\}$ 可以是与 DFT 相关的正交矢量组。这些矢量的计算公式为：$\boldsymbol{v}_i{}'=k^{-1/2}$（1，$\mathrm{e}^{-\mathrm{j}\frac{2\pi}{k}(i-1)}$，…，$\mathrm{e}^{-\mathrm{j}\frac{2\pi}{k}(i-1)(k-1)}$）。在此，为了使残余噪声功率谱密度为对称函数，我们必须选择 $\alpha_i=\alpha_{k-i+2}$，$i=2$，…，$k/2$ 并假设 k 是偶数。这就意味着最

多只能使用 $k/2+1$ 个约束条件，DFT 相关的矢量组 $\{\boldsymbol{v}_i\}$ 可以使用与残余噪声听觉相符的约束条件。

为了实现最佳的滤波器，假设 $\boldsymbol{e}_l$表示 $\Re^k$空间中的单位矢量，其中第 l 个分量为 1，其他所有分量为 0。将 $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m\}$ 扩展成一个 $k \times k$ 的正交矩阵或西矩阵 $\boldsymbol{V}=\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\}$，令 $\mu_i=0$（其中 $m<i\leqslant k$），$\boldsymbol{M}=\mathrm{diag}(k\mu_1, \cdots, k\mu_k)$ 表示 k 倍的拉格朗日乘子矩阵，该矩阵是非负的。令 $\boldsymbol{Q}=\boldsymbol{R}_w^{-1/2}\boldsymbol{U}$，$\boldsymbol{T}=\boldsymbol{Q}'\boldsymbol{U}$。假定 $\boldsymbol{H}=\boldsymbol{H}_2$，最佳估计矩阵的计算公式如下[11]：

$$\boldsymbol{H}_2=\boldsymbol{R}_w^{1/2}\boldsymbol{U}\tilde{\boldsymbol{H}}_2\boldsymbol{U}'\boldsymbol{R}_w^{-1/2} \tag{20-11}$$

式中，$\tilde{\boldsymbol{H}}_2$的列计算公式为

$$\tilde{\boldsymbol{h}}_l=\boldsymbol{T}\lambda_l(\boldsymbol{M}+\lambda_l\boldsymbol{I})^{-1}\boldsymbol{T}^{-1}\boldsymbol{e}_l, l=1,\cdots,k \tag{20-12}$$

对于所有的 i 和 l，$k\mu_i \neq -\lambda_l$。最佳估计算法首先对噪声进行白化处理，然后利用从白化信号的协方差矩阵的特征分解中得到正交变换 $\boldsymbol{U}'$，结合矩阵$\tilde{\boldsymbol{H}}_2$对最后的元素进行修正，这个过程和式（20-6）中的估计算法非常相似。但是，如果输入噪声是有色噪声的话，那么矩阵$\tilde{\boldsymbol{H}}_2$就不是对角矩阵了。

如果噪声是白色噪声，其方差为 σ_w^2，而且 $\boldsymbol{V}=\boldsymbol{U}$，$m=k$，式（20-10）的优化问题就会变得很简单，因为已知的输入和输出噪声方差惟一地确定了过滤矩阵 $\boldsymbol{H}$，该过滤矩阵的计算公式为 $\boldsymbol{H}=\boldsymbol{U}\boldsymbol{G}\boldsymbol{U}'$，其中 $\boldsymbol{G}=\mathrm{diag}(\sqrt{\alpha_1}, \cdots, \sqrt{\alpha_k})$[7]。对于这种情况，启发式因子

$$\alpha_i=\exp\{-v\sigma_w^2/\lambda_i\}(\text{其中}, v\geqslant 1) \tag{20-13}$$

作为一个训练常数，在实际应用中非常有效。这个因子是在观察 $v=2$ 时发现的；在观察过程中，$\alpha_i{}^{-1/2}$的一阶泰勒展开式可以生成一个估计算法矩阵 $\boldsymbol{H}=\boldsymbol{U}\boldsymbol{G}\boldsymbol{U}'$，该矩阵与式（20-6）中的维纳（Wiener）估计矩阵是一致的，其中 $\mu=1$。在实际应用中，采用式（20-13）的估计矩阵比维纳（Wiener）滤波器的性能要好得多。

20.3　短期频谱估计

在另一个早期的语音增强应用中，清晰信号的短期频谱数量是从噪声信号中估计出来的。语音信号和噪声过程被假设为统计上的不相关，每个过程的频谱分量都为零平均值，在统计上与高斯随机变量不相关。假设 $A_y\mathrm{e}^{\mathrm{j}\theta_y}$表示一个给定帧中清晰信号 Y 的频谱分量；$A_z\mathrm{e}^{\mathrm{j}\theta_z}$表示一个对应噪声信号的频谱分量；$\sigma_y^2=E\{A_y^2\}$和 $\sigma_z^2=E\{A_z^2\}$ 分别表示清晰信号和噪声频谱分量的方差。如果给定帧

中对应的噪声频谱分量的方差是由σ_w^2表示，那么我们可以得到$\sigma_z^2=\sigma_y^2+\sigma_w^2$。令

$$\xi=\frac{\sigma_y^2}{\sigma_w^2};\ \gamma=\frac{A_z^2}{\sigma_w^2};\ \vartheta=\frac{\xi}{\xi+1}\gamma \tag{20-14}$$

根据$A_z e^{j\theta_z}$，A_y的 MMSE 估计算法如下[4]：

$$\hat{A}_y=\Gamma(1.5)\frac{\sqrt{\vartheta}}{\gamma}\exp\left(-\frac{\vartheta}{2}\right)\left[(1+\vartheta)I_0\left(\frac{\vartheta}{2}\right)+\vartheta I_1\left(\frac{\vartheta}{2}\right)\right]A_z \tag{20-15}$$

式中，$\Gamma\ (1.5)=\frac{\sqrt{\pi}}{2}$；$I_0(.)$和$I_1(.)$分别表示修正后的零阶和一阶贝塞尔（Bessel）函数。类似于式（20-5）中的维纳（Wiener）滤波器，该估计算子也要求已知每个信号和噪声频谱分量（分别为σ_y^2和σ_w^2）的二阶统计特性。

为了得到清晰信号频谱分量的估计算法，我们将结合式（20-15）中的频谱数量估计算子与该分量的复杂指数估计算子一起计算。假设$e^{j\hat{\theta}_y}$表示$e^{j\theta_y}$的估计算子，该估计算子是噪声频谱分量$A_z e^{j\theta_z}$的函数。复杂指数$e^{j\theta_y}$的 MMSE 估计算子可以从下面的公式中得到：

$$\min_{e^{j\hat{\theta}_y}} E\{|e^{j\theta_y}-e^{j\hat{\theta}_y}|^2\} \qquad \text{条件为：}|e^{j\hat{\theta}_y}|=1 \tag{20-16}$$

式（20-16）中的约束条件确保了当两者结合计算时，估计算子$e^{j\hat{\theta}_y}$不会影响到估计算子$\hat{A}_y$的最优性。式（20-16）中受到约束条件限制的最小化问题使得$e^{j\hat{\theta}_y}=e^{j\theta_z}$，这样就简化了噪声信号的复杂指数。

注意到，式（20-8）中的维纳（Wiener）滤波器具有零相位，因此它可以有效使用噪声信号的复杂指数$e^{j\theta_z}$来估计清晰信号的频谱分量。因此，式(20-8)中的维纳（Wiener）估计算子和式（20-15）中的 MMSE 估计算子都可以使用噪声相位的复杂指数。但是，通过维纳（Wiener）滤波器得到的清晰信号的频谱数量估计，不是 MMSE 中最佳的。

估计A_y的其他标准也可以使用。例如，A_y可以从下面的公式进行估计：

$$\min_{\hat{A}_y} E\{(\lg A_y-\lg\hat{A}_y)^2\} \tag{20-17}$$

该标准的目的是产生一个A_y的估计，其对数尽可能接近 MMSE 中A_y的对数。这个感性上的标准得到的结果如下所示[5]：

$$\hat{A}_y=\frac{\sigma_y^2}{\sigma_y^2+\sigma_w^2}\exp\left(\frac{1}{2}\int_{\vartheta}^{\infty}\frac{e^{-t}}{t}dt\right)A_z \tag{20-18}$$

式（20-18）中的积分就是著名的ϑ指数积分，该积分可以估计出数字值。

在另一个例子中，A_y^2可以从下面的公式估计得到：

$$\min_{\hat{A}_y} E\{(A_y^2 - \hat{A}_y^2)^2\} \tag{20-19}$$

而且，A_y的估计可以从$\sqrt{\hat{A}_y^2}$得到。式（20-19）中标准的目的是对 MMSE 中清晰信号频谱分量的数量平方进行估计。该估计算子在后面增强信号的处理中非常有用，例如，对低传输速率信号编码的自回归分析应用[13]，在该应用中，清晰信号的自回归函数估计算子可以从估计算子 A_y^2中得到。式（20-19）对应的最佳估计算子是一个非常有名的算子，如下所示（例如参考文献［6］）：

$$\hat{A}_y^2 = \frac{\sigma_y^2 \sigma_w^2}{\sigma_y^2 + \sigma_w^2} + \left| \frac{\sigma_y^2}{\sigma_y^2 + \sigma_w^2} A_z \right|^2 \tag{20-20}$$

20.4 多状态语音模型

20.2 节和 20.3 节中介绍的所有估计算子给出了一个前提条件，即语音信号始终处于噪声信号中。在 20.2 节的符号中，$\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{W}$。由于帧的长度相对来说是很短的（为 30 ~ 50ms 级别），因此，我们更适合假设包含噪声的语音信号概率为 η，而不含噪声的语音信号概率为 $1-\eta$。这样，我们就有了两种假设，一种包含噪声的语音信号（假设为 $\boldsymbol{H}_1$），另一种不含噪声的语音信号（假设为 $\boldsymbol{H}_0$），它们的概率分别为 η 和 $1-\eta$。而且，

$$\boldsymbol{Z} = \begin{cases} \boldsymbol{Y} + \boldsymbol{W} & \boldsymbol{H}_1\text{时} \\ \boldsymbol{W} & \boldsymbol{H}_0\text{时} \end{cases} \tag{20-21}$$

在不确定模型情况下，A_y的 MMSE 估计算子可以表示为

$$\begin{aligned} \tilde{A}_y &= \Pr(\boldsymbol{H}_1 | \boldsymbol{Z}) \cdot E\{A_y | \boldsymbol{Z}, \boldsymbol{H}_1\} + \Pr(\boldsymbol{H}_0 | \boldsymbol{Z}) \cdot E\{A_y | \boldsymbol{Z}, \boldsymbol{H}_0\} \\ &= \Pr(\boldsymbol{H}_1 | \boldsymbol{Z}) \cdot \hat{A}_y \end{aligned} \tag{20-22}$$

式中，$E\{A_y | \boldsymbol{Z}, \boldsymbol{H}_0\} = 0$，$E\{A_y | \boldsymbol{Z}, \boldsymbol{H}_1\} = \hat{A}_y$。

式（20-22）给出了更加实际的估计算子，而该公式在提高式（20-15）中估计算子的性能方面非常有用。该模型也可以得到其他的估计算子。注意到，式（20-21）中的模型不适合式（20-18）中的估计算子，因为式（20-17）中的 A_y 必须为正数才有意义。在不含噪声的语音不确定模型下的语音增强技术是由 McAulay 和 Malpass 在他们的开创性工作中首先提出来并使用的[16]。

上面这种模型的展开式假设语音矢量在不同的时刻处于不同的状态。不含噪声的语音不确定模型中假设两种状态分别代表了不含噪声的语音和包含噪声的语音。在另一种模型 Drucker 中[3]，有 5 种状态分别代表了摩擦音、元音、调整音

调声、滑音和鼻音，不含噪声的语音状态可以添加到该模型中并作为第6种状态。该模型的每个状态都需要一个如同式（20-22）中的估计算子。

如果采用隐藏马尔可夫过程（HMP）来模拟语音信号，就可以进一步扩展上面的模型（其中随时可以得到多个状态）了。HMP是一个二元随机状态和观察序列过程，状态过程 $\{S_t,\ t=1,\ 2,\ \cdots\}$ 是一个有限状态齐次马尔可夫链，对该过程无法直接进行观察。观察过程 $\{Y_t,\ t=1,\ 2,\ \cdots\}$ 是条件无关的状态过程。因此，每个观察过程在统计上都只取决于相同时间的马尔可夫链的状态，而与任何其他状态或观察过程无关。例如，假设一个HMP处于一个相加性白色噪声过程 $\{W_t,\ t=1,\ 2,\ \cdots\}$ 中，对于每个 t，令 $\boldsymbol{Z}_t=\boldsymbol{Y}_t+\boldsymbol{W}_t$ 表示噪声信号，$\boldsymbol{Z}^t=\{\boldsymbol{Z}_1,\ \cdots,\ \boldsymbol{Z}_t\}$；$J$ 表示马尔可夫链的状态数。给定 $\{\boldsymbol{Z}^t\}$ 条件下，$\boldsymbol{Y}_t$ 的因果MMSE估计算子为

$$\begin{aligned}\hat{Y}_t &= E\{\boldsymbol{Y}_t|\boldsymbol{Z}^t\}\\ &= \sum_{j=1}^{J}\Pr(S_t=j|\boldsymbol{Z}^t)E\{\boldsymbol{Y}_t|S_t=j,\boldsymbol{Z}_t\}\end{aligned}\qquad(20\text{-}23)$$

当 $J=2$，$\Pr(S_t=j|\boldsymbol{Z}^t)=\Pr(S_t=j|\boldsymbol{Z}_t)$ 时或者各状态之间在统计上不相关时，式（20-23）中的估计算子就可以简化成式（20-22）。

HMP是一个参数过程，该过程取决于马尔可夫链的初始分布和变换矩阵，以及给定状态下观察过程的条件分布参数。HMP的参数可以根据训练数据在离线条件下进行估计，之后可以用于构建式（20-23）中的估计算子。这种方法在理论上具有很强吸引力，因为这种方法提供了一种可靠的语音信号统计模型；这种方法在直观上也很有吸引力，因为语音信号会根据不同的本质特征区分成一串串的声音组，而且每一组都对应一个合适的滤波器。实现这种方法的难点是在将噪声语音的矢量映射到HMP的各个状态过程中时，如何实现较低的误差率。译码误差会导致语音矢量与预先设计的估计算子 $\{E\ \{\boldsymbol{Y}_t|S_t=j,\ \boldsymbol{Z}_t\},\ j=1,\ \cdots,\ J\}$ 之间出现关联错误，从而产生过滤成功率很低的语音矢量。另外，这种方法的复杂性也会随着状态数量的增加而增加，因为每个信号矢量都必须经过所有 J 个滤波器处理。

上面介绍的这种方法在应用时可以基于其他语音信号模型，例如，在参考文献［20］中，一个基于估计斜率周期的谐波过程可以用来模拟语音信号。

20.5 二阶统计估计

20.2节和20.4节中介绍的所有估计算子都基于清晰信号和噪声过程的某些统计特性，这些过程假设是一些已知的推理演绎过程。式（20-5）和式(20-11)中的信号子空间估计算子要求信号和噪声的协方差矩阵是已知的；式（20-15）、

式（20-18）和式（20-20）中的频谱数量估计算子要求语音信号以及噪声过程的每个频谱分量的方差是已知的。如果没有明确知道清晰信号和噪声过程的二阶统计特性，这些统计特性必须根据训练序列来估计，或者直接根据噪声来进行估计。

我们注意到，如果清晰信号和噪声过程的二阶统计特性估计值代替了给定估计方法中真正的二阶特性，那么该估计方法的最优性就得不到保证了。这些二阶统计特性的估计质量是语音增强系统整体性能中的关键。二阶统计特性的估计过程可以通过各种方法实现，这些方法通常可以参考频谱估计的理论[19]。本节重温了其中的部分方法。

根据训练数据来估计语音信号二阶统计特性的方法已经成功应用于清晰信号的编码和识别过程中了。在编码应用中，该方法是通过矢量量化来实现的；而在识别应用中，该方法是通过隐藏马尔可夫建模来实现的。但是，当只有噪声信号可供使用时，将一个给定的语音帧与矢量量化器的代码或者 HMP 的状态相匹配的过程很容易受到译码误差的影响。这些误差是语音信号真正的二阶统计特性和估计二阶统计特性之间出现不一致的根本原因，这种不一致在应用中会导致噪声语音信号帧使用错误的滤波器，而且会导致被处理的噪声信号产生让人无法接受的质量。

实践证明，根据一个噪声函数来对语音信号的二阶统计特性进行在线估计是一个不错的选择。由于分析帧长度相对较短，语音信号的协方差矩阵可以假设为 Toeplitz 矩阵。这样，必须估计每个分析帧中清晰信号的自相关函数。窗口自相关函数估计的傅里叶变换可以得到清晰信号频谱分量方差的估计。

对于一个带宽敏感的稳定噪声过程来说，噪声自相关函数可以根据噪声信号的原始分段进行估计，该噪声信号中只包含噪声。如果噪声过程不是一个带宽敏感的稳定过程，那么噪声信号帧必须归为式（20-19）中的一类，而且当检测到一个新的噪声帧时，必须更新噪声过程的自相关函数。

在信号子空间方法中[7]，首先根据窗口自相关函数的估计值对噪声和噪声过程的功率谱密度进行估计；然后，利用式（20-9）对清晰信号的功率谱密度进行估计，该估计过程是一个傅里叶反变换，可以得到一个清晰信号理想自相关函数的估计。在实现式（20-15）中的 MMSE 频谱估计算子时，清晰信号中每个频谱分量的方差（参考文献［4］中有介绍）通常使用一个递归估计算子（如参考文献［1，2］）。假设$\hat{A}_y(t)$表示 t 时刻语音帧中清晰信号频谱分量的数量估计算子，该估计算子可以是式（20-15）中的 MMSE 估计算子；$A_z(t)$ 表示噪声信号对应的频谱分量数量；$\hat{\sigma}_w^2(t)$ 表示该帧中噪声过程频谱分量的估计方差；$\hat{\sigma}_y^2(t)$ 表示该帧中清晰信号频谱分量的估计方差。该估计算子如下式所示：

$$\hat{\sigma}_y^2(t)=\beta \hat{A}_y^2(t-1)+(1-\beta)\max\{A_z^2(t)-\sigma_w^2(t),0\} \tag{20-24}$$

式中，$0\leqslant\beta\leqslant1$ 是一个实验常数。

我们在过去发现，该估计因子在实践中非常有效，它是一个启发式激发因子，而它的解析性能至今仍然是一个未知数。

根据给定的噪声信号采样函数来估计语音信号功率谱密度的参数法是由 Musicus 和 Lim 提出来的[18]。假设给定语音帧中的清晰信号是一个零平均值高斯自回归过程；噪声是一个零平均值白高斯过程，该过程与语音信号不相关。噪声信号的参数包括自回归系数、自回归过程增益以及噪声过程方差。这些参数可以利用 EM 算法在 ML 中进行估计；为实现这个过程，Musicus 首先提出了参数收敛性的 EM 算法和条件[17]。这种方法开始时对参数进行初始估计；该初始估计值用来计算清晰信号的条件平均估计值和给定噪声信号下的清晰信号采样方差矩阵的条件平均估计值；然后，这些估计值用来产生一个新的参数。上面的过程反复进行直到达到一个固定点或满足一个停止标准。

这种 EM 流程总结如下：假设一个梯状自回归过程 $\{Y_t,\ t=1,\ 2,\ \cdots\}$，其阶数为 r，系数为 $a=(a_1,\ \cdots,\ a_r)'$。假设 $\{V_t\}$ 表示零平均值单位方差高斯随机变量的独立均匀分布（iid）序列；σ_v 表示增益因子。自回归过程的采样函数如下所示：

$$y_t=-\sum_{i=1}^{r}a_i y_{t-i}+\sigma_v v_t \tag{20-25}$$

假设式（20-25）的初始条件全为 0，例如，$t<0$ 时，$y_t=0$。假设 $\{W_t,\ t=1,\ 2,\ \cdots\}$ 表示噪声过程，该过程包含一个零平均值单位方差高斯随机变量的 iid 序列。假设噪声自回归过程是一个 k 维矢量；$\boldsymbol{Y}=(Y_1,\ \cdots,\ Y_k)'$同样还定义了 $\boldsymbol{V}$ 和 $\boldsymbol{W}$，$\boldsymbol{V}$ 和 $\boldsymbol{W}$ 分别对应了过程 $\{V_t\}$ 和 $\{W_t\}$；$\boldsymbol{A}$ 表示低阶三角 Toeplitz 矩阵，其第一列为 $(1,\ a_1,\ \cdots,\ a_r,\ 0,\ \cdots,\ 0)'$。这样，

$$\boldsymbol{Z}=\sigma_v\boldsymbol{A}^{-1}\boldsymbol{V}+\sigma_w\boldsymbol{W} \tag{20-26}$$

假设 $\phi_m=(a(m),\sigma_v(m),\sigma_w(m))$ 表示第 m 次迭代之后 $\boldsymbol{Z}$ 的参数估计值；$\boldsymbol{A}(m)$ 表示由矢量 $\boldsymbol{a}(m)$ 构成的矩阵 $\boldsymbol{A}$；$\boldsymbol{R}_y(m)=\sigma_v^2(m)\boldsymbol{A}(m)^{-1}(\boldsymbol{A}(m)^{-1})'$表示从 ϕ_m 得到的式（20-25）中自回归过程的协方差矩阵；$\boldsymbol{R}_z(m)=\boldsymbol{R}_y(m)+\sigma_w^2(m)\boldsymbol{I}$表示基于 ϕ_m 的式（20-26）中噪声信号的协方差矩阵；$\hat{\boldsymbol{Y}}$和$\hat{\boldsymbol{R}}$分别表示清晰信号 $\boldsymbol{Y}$ 的条件平均估计值和给定 $\boldsymbol{Z}$ 时 $\boldsymbol{Y}$ 与当前估计 ϕ_m 的采样方差的条件平均估计值。在高斯假设条件下，我们可以得到：

$$\begin{aligned}\hat{\boldsymbol{Y}}&=E\{\boldsymbol{Y}|\boldsymbol{Z};\phi_m\}\\&=\boldsymbol{R}_y(m)\boldsymbol{R}_z^{-1}(m)\boldsymbol{Z}\end{aligned} \tag{20-27}$$

$$\hat{R} = E\{YY' | Z;\phi_m\}$$

$$= R_y(m) - R_y(m)R_z^{-1}(m)R_y(m) + \hat{Y}\hat{Y}' \quad (20\text{-}28)$$

式（20-27）和式（20-28）中清晰信号的统计特性估计过程是由 EM 算法的 E 算法构成的。

定义一个 $r+1 \times r+1$ 阶协方差矩阵 $\boldsymbol{S}$，其主体为：

$$S(i,j) = \sum_{l=\max(i,j)}^{k-1} \hat{R}(l-i,l-j), \quad i,j=0,\cdots,r \quad (20\text{-}29)$$

定义 $r \times 1$ 阶矢量 $\boldsymbol{q} = (S(1,0),\cdots,S(r,0))'$和 $r \times r$ 阶协方差矩阵 $\boldsymbol{Q} = \{S(i,j),i,j=1,\cdots,r\}$。在第 $m+1$ 次迭代之后，参数 ϕ 新的估计值为。

$$a(m+1) = -Q^{-1}q \quad (20\text{-}30)$$

$$\sigma_v^2(m+1) = (S(0,0) + 2a'(m)q + a'(m)Qa(m))/k \quad (20\text{-}31)$$

$$\sigma_w^2(m+1) = \mathrm{tr}(E\{(Z-Y)(Z-Y)' | Z,\phi_m\})/k$$

$$= \mathrm{tr}(ZZ' - 2\hat{Y}Z' + \hat{R})/k \quad (20\text{-}32)$$

式（20-30）和式（20-32）中噪声信号参数的计算过程由 EM 算法的 M 算法构成的。

注意到，增强信号作为该规则的副产品可以通过式（20-27）中的估计算子在最后一次迭代后得到。上面状态空间中的参数估计问题的公式表示法（假设有色噪声是另一种自回归过程）与式（20-27）和式（20-28）中的估计算子是通过卡尔曼（Kalman）过滤算法和滤波算法实现的。

上面介绍的 EM 法与之前介绍的由 Lim 和 Oppenheim 提出的方法是不同的[12]。它们之间进行比较时假设噪声方差是已知的。在 EM 法中，目标是获得自回归过程真实参数的 ML 估计值，这个目标是通过对过程参数组中的噪声信号的似然函数局部极大化来实现的。EM 法在最后可以得到清晰信号 $\boldsymbol{Y}$ 及其采样相关性矩阵 $\boldsymbol{YY'}$的迭代估计值，以及自回归过程（a，σ_v）参数的迭代估计值。参考文献［12］中介绍的方法其目标是对自回归过程的参数进行估计，该过程提取清晰信号和噪声信号的联合似然函数的最大值；通过对联合似然函数的迭代极大化，该方法最后可以得到清晰信号 $\boldsymbol{Y}$ 的迭代估计值，以及自回归过程（a，σ_v）参数的迭代估计值。因此，这两种方法之间的主要区别在于估计统计特性上，因为在每次迭代过程中，自回归参数是根据估计统计特性来进行估计的。这种区别会影响到参考文献［12］中规则的收敛性，该规则可以产生一个不一致的参数估计值；该规则比 EM 法更容易实现，因此它在部分作者中很受欢迎；为

了克服这种不一致性，他们在每一次迭代过程中开发了一组参数估计的约束条件。

20.6 结束语

到目前为止，我们已经了解了语音增强问题的很多方面，并介绍了几种艺术级的解决方法。特别是，我们弄清楚了语音增强技术中的各个重难点，并介绍了常用于清晰信号设计估计的统计模型和失真衡量标准。本章中，我们主要集中讨论了包含相加性互不相关宽带噪声的语音信号。如前所述（即使在那样情况下），一个可接受的通用解决方法是无效的，因而需要更多的研究来精炼当前的各种方法或者重新开发一种新方法。其他的噪声源（如室内回响噪音）带来了一个更大的挑战，因为噪声是一个不稳定的过程，该过程与信号直接相关，而且不容易模拟。因此，在未来，语音增强问题需要得到更多的关注和研究，以满足各种潜在应用和未来计算器件应用的要求。

名词解释

语音增强：用来提高给定语音信号采样的感觉方面的行为。

质量：一种感知语音信号的客观衡量标准。

可理解性：一种表示给定文本中文字所占比例的主观衡量标准，该文本期望能被收听者正确理解。

信号估计算子：观察噪声信号的函数，该噪声信号非常接近清晰信号。

期望-极大化：一种迭代方法，利用迭代估计和极大化过程来对参数进行估计。

自回归过程：一种随机过程，通过将白色噪声输入一个全极值点滤波器得到。

带宽敏感稳定性：随机过程的一种属性，该随机过程的二阶统计特性不会随时间变化。

渐近弱稳定性：随机过程的一种属性，用来表示可能的带宽敏感稳定性。

隐藏马尔可夫过程：通过噪声渠道来观察的一种马尔可夫链。

参考文献

[1] Cappé, O., Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor, *IEEE Trans. Speech and Audio Proc.*, vol. 2, pp. 345-349, Apr. 1994.

[2] Cohen, I. and Berdugo, B. H., Speech enhancement for non-stationary noise environments, *Signal Pro-*

cessing, vol. 81, pp. 2403-2418, 2001.

[3] Drucker, H., Speech processing in a high ambient noise environment, *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 165-168, Jun. 1968.

[4] Ephraim, Y. and Malah, D., Speech enhancement using a minimum mean square error short time spectral amplitude estimator, *IEEE Trans. Acoust.*, *Speech*, *Signal Processing*, vol. ASSP-32, pp. 1109-1121, Dec. 1984.

[5] Ephraim, Y. and Malah, D., Speech enhancement using a minimum mean square error Log-spectral amplitude estimator, *IEEE Trans. Acoust.*, *Speech*, *Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.

[6] Ephraim, Y., Statistical model based speech enhancement systems, *Proc. IEEE* vol. 80, pp. 1526-1555, Oct. 1992.

[7] Ephraim, Y. and Van Trees, H. L., A signal subspace approach for speech enhancement, *IEEE Trans. Speech and Audio Proc.*, vol. 3, pp. 251-266, July. 1995.

[8] Ephraim, Y. and Merhav, N., Hidden Markov Processes, *IEEE Trans. Inform. Theory*, vol. 48, pp. 1518-1569, June 2002.

[9] Gannot, S., Burshtein, D., and Weinstein, E., Iterative and sequential Kalman filter-based speech enhancement algorithms, *IEEE Trans. Speech and Audio Proc.*, vol. 6, pp. 373-385, July 1998.

[10] Gray, R. M., *Toeplitz and Circulant Matrices*: II. Stanford Electron. Lab., Tech. Rep. 6504-1, Apr. 1977.

[11] Lev-Ari, H. and Ephraim, Y., Extension of the signal subspace speech enhancement approach to colored noise, *IEEE Sig. Proc. Let.*, vol. 10, pp. 104-106, Apr. 2003.

[12] Lim, J. S. and Oppenheim, A. V., All-pole modeling of degraded speech, *IEEE Trans. Acoust.*, *Speech*, *Signal Processing*, vol. ASSP-26, pp. 197-210, June 1978.

[13] Lim, J. S. and Oppenheim, A. V., Enhancement and bandwidth compression of noisy speech, *Proc. IEEE* vol. 67, pp. 1586-1604, Dec. 1979.

[14] Lim, J. S., ed., *Speech Enhancement*. Prentice-Hall, New Jersey, 1983.

[15] Makhoul, J., Crystal, T. H., Green, D. M., Hogan, D., McAulay, R. J., Pisoni, D. B., Sorkin, R. D., and Stockham, T. G., *Removal of Noise From Noise-Degraded Speech Signals*. Panel on removal of noise from a speech/noise National Research Council, National Academy Press, Washington, D. C., 1989.

[16] McAulay, R. J. and Malpass, M. L., Speech enhancement using a soft-decision noise suppression filter, *IEEE Trans. Acoust.*, *Speech*, *Signal Processing*, ASSP-28, pp. 137-145, Apr 1980.

[17] Musicus, B. R., *An Iterative Technique for Maximum Likelihood Parameter Estimation on Noise Data*. Thesis, S. M., M. I. T, Cambridge, MA, 1979.

[18] Musicus, B. R., and Lim, J. S., Maximum likelihood parameter estimation of noisy data, *Proc. IEEE Int. Conf. on Acoust.*, *Speech*, *Signal Processing*, pp. 224-227, 1979.

[19] Priestley, M. B., *Spectral Analysis and Time Series*, Academic Press, London, 1989.

[20] Quatieri, T. F. and McAulay, R. J., Noise reduction using a soft-decision sin-wave vector quantizer, *IEEE Int. Conf. on Acoust.*, *Speech*, *Signal Processing*, pp. 821-824, 1990.

第 21 章　Ad Hoc 网络

Michel D. Yacoub　Paulo Cardieri　Élvio João Leonardo
Álvaro Augusto Machado　Medeiros

21.1　引言

Ad Hoc 网络是一种无线网络，该网络无需基本设施的辅助，也不用进行集中管理。Ad Hoc 网络由一组无线终端（节点）构成，任何两个终端之间采用存储延时机制进行通信。一个需要发送数据的终端首先必须接入媒质并将信息发送到邻近的其他终端，相邻的终端接收到信息后确认该信息是否就是访问该终端的信息；如果不是，这个终端就会存储该信息以便在一个合适的时间延时转发给另一个终端。这个过程持续进行直至转发到目的终端。我们可以注意到，在 Ad Hoc 网络中没有固定的路由器。各个节点都可以是移动的，而且它们之间可以以任意的方式进行动态互连。节点的功能就如同一个路由器，用来寻找并维持与网络中其他节点之间的路由。Ad Hoc 网络广泛应用于紧急事件和救护、会议、野外数据采集、传感器网络以及家庭和办公网络中。廉价的硬件、微型的接收装置、快速的处理器推动了无线 Ad Hoc 网络的快速发展。本章将着重从以下 4 个方面来介绍 Ad Hoc 网络：路由、媒体接入、TCP/IP 要点、容量。在路由方面，我们将详细介绍主要的路由算法；在媒体接入方面，我们将描述主要的媒体接入协议；在 TCP/IP 要点方面，我们将讨论 Ad Hoc 网络中与 TCP/IP 性能相关的各个方面；而在容量方面；我们将介绍一些与网络容量有关的公式表示法。

21.2　路由算法

Ad Hoc 网络中路由算法的设计是一项富有挑战性的任务。路由算法必须能提供高度的完善性和智能，以便有效处理无线系统中的有限资源；路由算法必须很强健，以适应各种恶劣的无线环境。同时，路由算法还必须具有很强的灵活性，以适应不断变化的网络环境，如网络规模、流量分布以及移动性。路由算法已经在无线通信系统中应用很久了，它们通常可以分为两大类：远距离矢量（Distant Vector，DV）算法和链接状态（Link-State，LS）算法。DV 算法可以为每个节点提供一个矢量，该矢量包含了该节点与所有目的节点之间的跳数以及下

一个节点与所有目的节点之间的跳数（一个节点与相邻节点之间的距离称为“一跳”）；LS 算法通过相邻节点的链接信息周期散布法为每个节点提供一个最新的网络拓扑图。这两个算法在无线和移动环境中的直接应用非常复杂。DV 协议经常会遇到慢性路由汇聚，并可能产生回路；而 LS 协议要求频繁使用资源来维持节点的更新，这样会产生很高的带宽负荷。

随着人们对无线网络的日益关注，各种新的路由算法逐渐克服 DV 协议和 LS 协议中的各种限制。这些新的路由算法可以分为 3 大类：主动的或者表格驱动、被动的或者按要求选择、混合的。主动协议要求节点一直保存路由信息表，该路由信息表中的信息只有在定期更新时或者网络拓扑发生变化时才会发生变化。这些算法根据保存的信息类型和更新的方式来进行区分。被动协议根据需求来产生路由，并通过一个路由发现过程来实现的，当发现一个路由或者所有可能的路由都被检查过之后，该过程就算完成。由散布法实现的路由发现过程必须通过网络来请求数据包。在确定一条路由之后，该路由通过路由保持流程来保持，直至从源节点到目的节点的每一条路径都不可接入，或者直到所有的路由都不符合要求。混合协议既是主动的也是被动的，各个距离很近的节点构成一个框架，在该框架中使用主动协议；距离较远的节点使用被动协议。

1. 主动算法

CGSR——簇头网关交换路由算法（Chiang，1997），在 CGSR 算法中，整个网络被分割成各个节点簇；在每一簇节点中，选择一个节点作为簇头。一个节点可能属于一个簇，这样该节点就是一个内部节点；或者属于多个簇，这样该节点就变成了一个网关。数据包可以从节点传输到簇头，也可以从簇头传输到节点。在这样的网络中，路由过程可以描述如下：假设源节点和目的节点属于不同的簇，源节点发送数据包至簇头，该簇头延时传输这些数据包至网关，网关再延时传输数据包至另一个簇头，这个过程一直继续下去直至目的节点。

DREAM——移动远距离效应路由算法（Basagni，1998）。在 DREAM 中使用了 GPS，以便每个节点都可以保存局部路由表，同时记录所有节点的位置。各个节点通过广播控制数据包来更新位置信息。一个需要发送数据包的源节点首先必须计算目的节点的方向，然后在各个方向选择一组只有一个跳数的相邻节点（如果不存在这样的相邻节点，那么数据就会溢出网络）。当数据帧头装入到各个方向的节点后，就可以进行数据传输了。那些在数据帧头中指定的节点才有资格接收和处理数据。路径上的所有节点反复进行上面的过程，直至目的节点。接收到数据包后，目的节点会采用相同的路由算法向源节点发送一个 ACK。

DSDV——目的地连续远距离矢量路由算法（Perkins，1994）。在 DSDV 算法中，每个节点保存一个路由表，其中包含了网络中的所有可能的目的节点以及每个目的节点的跳数。网络的入口是根据目的节点以连续的数字进行编号，这样移

动节点就可以区别新的路由和稳定的路由，以避免产生路由回路。路由表的连贯性是通过网络中的定期更新来保证的。

FSLS——模糊可见链接状态算法（Santivanez，2001）。在FSLS算法中，使用了最佳的链接状态更新算法（模糊可见链接状态）。路由每隔2^kT时间更新一次，其中k是跳数，T是最小链接状态更新传输周期；2^k表示等待更新的节点数。FSLS的工作原理与FSR非常相似，如下所述。

FSR——鱼眼状态路由算法（Iwata，1999；Pei，2000）。在FSR算法中，每个节点保存一个拓扑图，链接状态信息在相邻的节点之间定期进行交换；交换的频繁性取决于相邻节点与当前节点之间的跳数。邻近的目的节点更新得更加频繁，而较远的目的节点则相对较少更新。因此，FSR算法可以产生关于直接相邻节点的精确距离和路径信息，以及关于远距离节点不太精确的信息。不过，这些不太精确的信息在数据包到达目的节点后会得到补偿。

GSR——整体状态路由算法（Chen，1998）。在GSR算法中，基于相邻节点更新消息的链接状态表会定期交换链接状态信息。更新消息的大小会随着网络规模的增加而增加，在这种情况下，网络就需要一个很高的带宽了。

HSR——分级状态路由算法（Pei，1999）。在HSR算法中，链接状态原则和分级寻址法以及拓扑图结合一起使用。簇集算法可以用来将相邻的节点组织到各个簇中。每个节点都具有惟一的特性，这些特性包括MAC地址或分级信息。任意两个节点之间的通信都是通过物理连接或逻辑链接的方式进行的。物理连接支持节点之间的真实通信，而逻辑链接用于实现通信的分级结构，用来构建多级结构；其中最低的分级通常是物理层，而最高的分级构成了逻辑层。之后，从最低层向高层再回到最底层的通信就开始了。

MMWN——移动无线网络中的多媒体支持算法（Kasera，1997）。在MMWN算法中，使用了簇集分解结构，每个簇都有两种节点：交换节点和终点。终点之间不进行通信，它只与交换节点进行通信。在一个簇中，选择一个交换节点作为位置管理点，该管理点负责位置更新和位置寻找。这意味着，相比传统路由表驱动的算法，MMWN算法总的路由开销会大幅度下降。这样，MMWN中的信息就可以保存在动态分布的数据库中了。

OLSR——最佳链接状态路由算法（Jacquet，2001）。在OLSR算法中，每个节点通过定期变化的链接状态消息来保存网络拓扑信息。OLSR算法采用多点延时（MultiPoint Relays，MPR）策略来尽量减小控制消息的大小以及重广播节点的数量。为了实现这个目标，每个节点选择一组相邻的节点（多点延时——MPR）来重发数据包。那些没有被选中的节点可以读取和处理每个数据包，但是不能重发数据包。每个节点定期向它的一跳相邻节点进行广播。这些相邻节点中还可以挑选出一组能覆盖所有两跳相邻节点的一跳节点。每个目的节点的最佳

路由是在路由表中构建和保存的。

STAR——源树形自适应路由算法（Garci-Luna-Aceves，1999）。在STAR算法中，每个节点根据目的节点的参考路径保存一个源节点树形结构。该算法采用最低开销路由法（Least Overhead Routing Approach，LORA）来降低网络中的路由开销。消息在数量上的减少是通过对某些事件的发生设置最新传播条件来实现的。

TBRPF——基于反向路径的拓扑广播算法（Bellur，1999；Ogier，2002）。在TBRPF算法中，建立了两个独立的模型：相邻节点发现模型和路由模型。第一个模型负责执行区别性的HELLO消息，该消息中只包含了相邻节点的变化信息（加入或丢失）；第二个模型在部分拓扑信息的基础上进行工作，这些信息是通过定期的区别性拓扑更新得到的。如果一个节点 n 需要发送更新消息，网络中的每个节点就会朝着节点 n 的方向选择它们的下一个跳数节点（父节点）。链接状态更新信息会在生成树上相反的方向进行传播，该生成树是由所有节点至发送节点之间的最小跳数路径形成的。这就意味着节点 n 处发送的更新信息只有来自于对应的父节点才会被接受，这些更新信息之后会朝着节点 n 固有的子节点方向传播。

WRP——无线路由协议（Murthy，1996）。在WRP中，每个节点保存如下的一组路由表：距离表、路由表、链接成本表、消息重发目录（Message Retransmission List，MRL）表。MRL表的每个项目包含一串连续编号的更新消息、一个重发计数器、一个回执（每个相邻节点的回执都需要标记矢量，该矢量为一维矢量）以及一列在更新消息中发送的更新信息。MRL表记录了哪一个更新信息需要重发，哪一个相邻节点应该确认重发。更新消息只在相邻节点之间发送，而且只有在处理过更新消息之后或与相邻节点的链接发生变化时才会产生更新消息。各个节点可以通过接收ACK消息和其它消息来了解相邻的节点信息。如果一个节点不需要发送任何类型的消息，那么它将会在特定的时间段内发送一个HELLO消息来确认链接在线。

2. 被动算法

ABR——基于关联性的路由算法（ABR）（Toh，1996）。在ABR中，采用查询延时技术来确定源节点至目的节点之间的路由，并根据关联性标记来选择稳定的路由，该关联性标记由每个节点与其相邻的节点以及选定的更高关联性标记链接共同保持。这样，最短的路径就无法实现，取而代之的是较长的路径。在这种情况下，网络中就需要更少的路由重构，从而留下了更多的带宽可用。ABR算法需要周期信标来确定链接的相关性，该链接要求所有的节点始终保持活动，这样会产生额外的功耗。

AODV——Ad Hoc需求距离矢量算法（Das，2002）。在AODV算法中，使

用了周期信标和连续编号流程。数据包只发送目的地址，而不会发送完整的路由信息，这种情况在路由应答中也会出现。AODV 算法的优点是它可以适应高动态的网络。换句话说，各个节点在路由构建中会经历很长的延时。

ARA——基于"蚁群"的路由算法（Günes，2002）。在 ARA 中，利用了蚂蚁寻找食物的行为原理来降低路由开销。在寻找食物时，蚂蚁们为后面的蚂蚁留下一些踪迹（外激素），后来的蚂蚁们根据这些踪迹来寻找食物。在路由发现过程中，ARA 在网络中传播一个前向 ANT（FANT）直至目的节点；然后，返回一个后向 ANT（BANT），这样就形成了一条路径，数据包传输就可以开始了。路由是通过增加或减小每个节点"外激素值"来保持的。每当一个数据包通过给定节点时，该节点的"外激素值"就会增加；没有数据包通过时，该"外激素值"就会减小直至耗尽。我们可以推断出，FANT 和 BANT 的大小很小，因此，每个控制数据包的路由开销就会很小。

CBRP——基于簇的路由协议（Jiang，1999）。在 CBRP 中，各个节点被划分为各个簇，一个簇包含一个簇头。与传统的散布法相比，分级法可以减少网络的控制开销。当然，分级法还会产生相关的簇信息开销和簇保持开销。分级结构产生的较长传输延时会给节点带来不一致的拓扑信息，从而导致临时性的路由回路。

DSR——动态源路由算法（Johnson，2002）。在 DSR 算法中，不存在周期信标（HEELO 消息），该周期信标是一项重要特征，节点在接收到该信标后会进入休眠状态，从而可以节省电池电量。DSR 中的每个数据包发送完整的路由地址，这一点对于高移动性网络来说是一种缺陷。另一方面，各个节点在它们的路由缓冲区中可以保存多条路由，这一点带来的好处是节点在进行初始路由发现之前，可以检查它的路由缓冲区中是否存在有效路由；对于低移动性网络来说，这是一种优势特征。

FORP——面向数据流的路由协议（Su，1999）。在 FORP 中，由于移动性导致的路由故障可以通过以下的算法降到最低。一个 Flow-REQ 消息在整个网络中传播，接收到该消息的节点（基于 GPS 信息）根据前一个跳数来估算一个链接截止时间（Link Expiration Time，LET），并将其添加到 Flow-REQ 数据包中，数据包然后被重发。到达目的节点之后，每个节点会根据所有 LET 中的最小值来估算一个路由截止时间（RET）。其后，向源节点返回一个 Flow-SETUP 消息。这样，当产生链接故障时，仍然可以预见目的节点。如果遇到了链接故障，就会产生一个 Flow-HANDOFF 消息并以相同的方式进行传播。

LAR——位置辅助路由算法（Ko，1998）。在 LAR 算法中，利用位置信息可以使路由开销最小，这一点在传统的散布法中很常见。假设每个节点由 GPS 提供，数据包从一个跳数传输至另一个跳数时，将沿着与目的节点之间相对距离越

来越小的方向进行传输。

LMR——轻型移动路由算法（Corson，1995）。在LMR算法中，采用散布技术来确定所需的路由。在各个节点中，都保存了多条路由，这主要是出于可靠性的目的，以避免重新开始发现路由。另外，路由信息只涉及到相邻的节点，而不会涉及到整个路由。

RDMAR——相对距离微发现Ad Hoc路由算法（Aggelou，1999）。在RDMAR算法中，使用相对距离微发现流程来使路由开销最小，这个流程是通过计算源节点和目的节点之间的距离，并将每个路由请求数据包限制在一定的跳数内来实现的（例如，路由发现流程只限于局部区域）。事实上，这种算法只适用于源节点和目的节点之间之前的通信已经建立的情况。其他情况下，则使用散布法流程。

ROAM——非循环多路径需求路由算法（Raju，1999）。在ROAM算法中，使用了内部节点坐标和定向非循环子图，它们来自于路由器与目的节点之间的距离。如果无法达到理想的目的节点，那么多路径散布搜索就会停止。另外，每当路由器与目的节点之间的距离变化超过给定的门限值时，路由器就会向它的邻近节点广播更新消息。这种机制增强了网络的连接性，但代价是节点无法进入休眠状态来节省电池电量。

SSA——信号稳定自适应算法（Dube，1997）。在SSA算法中，路由选择是根据信号强度和位置稳定性来进行的，与关联性标记无关。另外，向目的节点发送的路由请求不能由中间的节点应答，这样在一个有效路由被发现之前就会产生延时，这主要是因为目的节点负责选择路由来进行数据传送。

TORA——临时有序路由算法（Park，1997）。在TORA中，设计要点是极个别接近拓扑变化事件的节点中控制消息的定位。各个节点保存了关于一跳节点的路由信息。路由生成和路由保持阶段采用一个高度机制来在目的节点建立定向非循环图。然后，链接就会根据与相邻节点之间的相对高度来确定是作为上行链接还是下行链接。路由删除阶段通过向网络中散发一个广播清晰数据包来删除无效路由。

3. 混合算法

DDR——分布式动态路由（Nikaein，2001）。在DDR算法中，使用了一个树形路由协议（简称“树”），但不要根节点。该树形协议是通过周期信标消息来建立的，这些消息只在相邻的节点之间进行交换。不同的树之间是通过网关节点相互连接的，从而构成了一片“森林”。每个树占据一块区域，并指配一个ID。路由是由混合Ad Hoc网络确定的。

DST——基于分布式生成树的路由算法（Radhakrishnan，1999）。在DST算法中，所有节点被划分成各个树，在每个树中，节点就变成了一个路由节点或一

个内部节点。树的根（也是一个节点）控制着树的结构。这种机制在 DST 算法中变成一种缺陷，因为根节点可能会生成另一个单一的故障点。

SLURP——可升级位置更新路由协议（Woo, 2001）。在 SLURP 算法中，节点被划分到各个互不重叠的区域，并为每个节点指定一个归属区域。每个节点的归属区域是通过静态映射函数来确定的，该函数对所有节点都是已知的，其输入信号为节点的 ID 和节点的数量。因此，所有节点都可以为每个节点确定归属区域。节点在归属区域中的当前位置是通过向它的归属区域单播一个位置更新数据包来维持的，一旦该更新数据包到达归属区域，该数据包就会广播给该区域中的所有节点。

ZHLS——基于区域的分级链接状态算法（Joa-Ng, 1999）。在 ZHLZ 算法中，采用分级拓扑并结合使用 GPS 来确定两种层结构：节点层和区域层。这样，每个节点都具有一个节点 ID 和一个区域 ID。如果一个节点需要一个到其他区域中的路由，那么源节点就会向所有其他区域广播一个区域层位置请求。相比被动协议中的散布法，这个过程产生的路由开销更低。节点在自己区域中的移动性可以保持网络的拓扑，这样就不需要进一步的位置搜索了。在 ZHLS 算法中，假设所有节点都包含一个预先编好的初始操作静态区域图。

ZRP——区域路由协议（Haas, 1999）。在 ZRP 中，一个路由区域在建立时定义了各跳之间的距离，该距离是网络连接的基础。这样，路由区域中的各个节点就可以立即获得有效路由。区域之外的节点路由是被动产生的（按需求产生），而且任何被动算法都可能使用。

21.3 媒体接入协议

无线 Ad Hoc 网络的媒体接入控制（Medium Access Control, MAC）是目前研究非常活跃的一个主题。网络的特性、各种不同的物理层有效技术以及各种预想服务的范围导致接入共享媒质的算法设计成为了一项非常艰难的任务，该共享媒质必须是高效、公平、功耗敏感和限制延时的。下面将介绍无线 MAC 协议与有线网络协议（Chandra, 2002）的区别要点。

1）半双工操作。由于自相接入的原因（例如，能量从发送端直接泄漏到接收端），我们很难构建可以同时接收和发送的终端。因此，在发送数据的同时进行冲突检测是不可能的，而且类似以太网的协议也不能使用。由于冲突无法检测，因此无线 MAC 协议采用冲突避免机制来尽量减小冲突发生的概率。

2）时变信道。在多路径衰退信道中，接收到的信号是发送信号的时移和消弱版本。由于信道特性和终端相对位置的变化，信号包络就变成了一个时变函数；而经过衰减信道的信号就更加严重了。各个节点要建立无线连接必须彻底认

识信道，以便评估通信链接状态。

3）突发信道错误。无线信道会比有线传输产生更高的误码率。除此之外，当信号衰减时，在突发数据包中也会产生误差，从而产生高概率的数据包丢失。因此，必须执行确认机制，以便在发生数据包丢失的情况下进行数据重传。

4）位置相关的载波侦听。由于信号强度会随距离增加而按照幂律衰减，因此只有在指定距离内的节点才能通信。这种现象导致了隐蔽终端和外露终端以及俘获遮蔽效应的出现。

5）隐蔽终端。图21-1给出了终端A、B和C的相对位置。其中，终端B处于终端A和C的覆盖范围之内，而终端A和终端C都相互处于各自的覆盖范围之外。如果终端A正在向终端B发送数据，而终端C也想要向终端B发送数据，那么终端C就会错误地检测出信道处于空闲状态，因为终端C处于终端A的覆盖范围之外，而终端A当前正处于通信繁忙状态。隐蔽终端问题可以在发送数据之前通过请求发送/清除发送（Request-To-Send/Clear-To-Send，RTS/CTS）握手协议（稍后将会介绍）来解决。

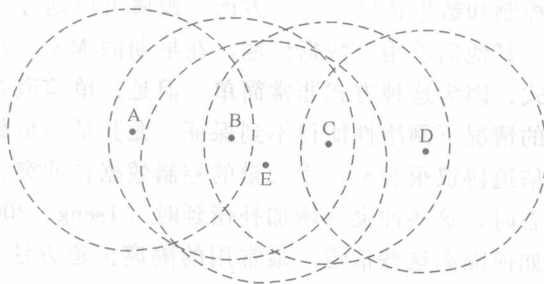


图21-1 隐蔽-外露终端问题

6）外露终端。外露终端是指那些处于发射点覆盖范围之内但处于接收点覆盖范围之外的终端。在图21-1中，如果终端B正在向终端A发送数据，终端C就会检测到终端B的信道处于繁忙状态。但是，由于终端C处于终端A的覆盖范围之外，因此终端C无法接入到当前的通信中去。这样，终端C可以利用信道来与其他终端建立一个并行链接，这些终端处于终端B的覆盖范围之外，如终端D。在这种情况下，终端C对于终端B就是一个外露终端。外露终端可能会导致信道处于次使用状态。如同隐蔽终端问题一样，这个问题也可以通过使用请求发送/清除发送（RTS/CTS）握手协议来解决。

7）俘获。给定终端处的俘获只有当多个信号同时到达该终端时才会发生，这些信号中有一个信号的信号强度超过了其他所有信号的信号强度和。在图21-1中，终端C和终端E都处于终端B的覆盖范围之内，如果终端C和终端E同时向终端B发送数据，那么在终端B处就会产生冲突。但是，如果其中一个信号强度远比另一个信号大，那么终端B就可以成功接收该信号，例如终端E的信号。俘获动作可以提高吞吐量，因为它降低了冲突发生的概率。但是，俘获动作更加倾向于那些离目的节点更近的发射点，这种现象会导致信道分配出现不

公平。

从上面的考虑，我们可以很快联想到，相比有线通信系统，Ad Hoc 网络 MAC 协议的设计必须进行充分考虑，因为 Ad Hoc 网络需要一组不同的参数。

1. 协议类型

在研究调查了大量的当前 MAC 协议后，Jurdak 等人（Jurdak，2004）提出了一组关键特征用来对 Ad Hoc 网络的 MAC 协议进行分类。

（1）信道隔离与接入协议

媒质的组织方式是协议设计中的重点。例如，所有站点可能共享一个信道来进行控制和数据传输。另一方面，媒质可以划分为多个信道，通常一个信道用于控制，其他信道用于数据传输。在早期的 MAC 设计中大多数都倾向于使用单信道方式，因为这种方式非常简单。但是，单信道方式容易产生冲突，而且在流量较大的情况下网络性能得不到保证。尤其是当负载较重时，通过网络仿真可以发现单信道协议很容易产生大量的控制数据包冲突，如 RTS 和 CTS，当媒质处于空闲状态时，这些冲突会增加补偿延时（Tseng，2002）。采用多信道时，我们必须知道如何隔离这些信道。最常用的隔离信道方法包括频分多址（Frequency Division Multiple Address，FDMA）、时分多址（Time Division Multiple Access，TDMA）和码分多址（Code Division Multiple Access，CDMA）技术。FDMA 技术采用多个载波来将媒质划分成多个频率间隙。FDMA 允许多个传输同时进行，每个发送端只能使用指定频率间隙的带宽。TDMA 技术将媒质划分成各个固定长度的时隙。多个时隙可以构成一个时间帧，并可以定义时隙的重复频率。由于周期特性的原因，TDMA 协议非常适合对延时敏感的流量的传输。在 TDMA 技术中，发送端可以使用一个时隙周期的指定完整可用带宽。另外，为了接入媒质，终端必须对帧和时隙非常清楚，因此，TDMA 协议要求各个终端必须是同步的。CDMA 技术允许发送端使用所有时间内的完整可用带宽。每个发送端都会被指定一串正交代码，这样用户就可以同时进行多个传输，并根据他们的惟一性代码来进行区分。CDMA 技术中的一个必要条件是功率控制，因为强度比理想信号大的干扰信号在接收端天线处可能会淹没掉理想信号，这就是著名的“远近效应（Near-Far)”。类似于 CDMA 技术，空分多址（Space Division Multiple Access，SDMA）技术允许发送端使用所有时间内的完整可用带宽。但是，终端必须使用定向天线，而且只有当理想传输方向不会干扰正在进行的通信时，才允许进行数据传输。

（2）RTS/CTS 握手协议

Ad Hoc 网络中很多 MAC 协议使用了不同的 RTS/CTS 握手技术。最初的三向握手技术可以减少隐蔽终端问题和外露终端问题的出现。一个需要发送数据的终端首先必须检测信道，如果信道中有很多合适的空闲时间，那么该终端就会发

送一个短的请求发送（RTS）数据包；所有侦听RTS数据包的终端都会延迟对它们的传输。然后目的终端就会返回一个清除发送（CTS）数据包，所有侦听CTS数据包的终端同样会延迟对它们的传输。接收到CTS的发送端就认为信道是理想的，并开始进行数据传输。

（3）拓扑结构

Ad Hoc网络具有很强的灵活性和不确定性；终端可能是移动的，而且拥有不同的性能和资源。Ad Hoc网络必须充分考虑到这些，并能在优化性能和减小功耗时动态适应各种变化（Jurdak，2004）。网络拓扑结构可以是集中式、簇式或者平台式。集中式拓扑结构中只有一个终端或基站用来控制和管理网络，该终端负责广播网络操作的相关信息。另外，所有终端只能通过中心终端进行通信。簇式拓扑结构生成一个局部的集中式网络，在该网络中某个终端负责承担中心终端的部分或所有职责。平台式拓扑结构实现了一种完全分布的方式，在该方式中所有终端处于相同的等级，不需要中心控制。平台式拓扑可以进一步划分成单跳和多跳结构。单跳结构假设目的节点处于发射点的覆盖范围之内；多跳结构假设目的节点可能超出了发射点可到达的相邻节点范围，如果是这样，中间的终端就负责中继数据包，直到到达目的节点。单跳协议更加简单，但是在网络规模上存在局限；多跳协议可以提高网络的规模，但是其复杂性更高。

（4）功耗协议

功耗是所有无线网络都比较关注的一项特征。功耗守恒定律对于移动终端来说是很重要的，因为移动终端电池的可用功耗是有限的。一个有效的功耗守恒策略涉及到很多方面。用来发射信号的能量占用了功耗中的很大一部分。理想情况是，所用的发射功耗应该刚好足够到达目的终端。另一种浪费能量的情况是终端需要长期检测信道或侦听相关通信。如果终端可以预先知道媒质何时处于繁忙状态，那么终端在其他时间就可以进入休眠状态，从而节省功耗。网络的性能可能会受到终端电池电量的影响，例如，在选择簇头或在指定传输优先级时，终端可以根据它的电池电量来相应地调整它的行为。在传输数据之前进行控制消息的交换也是浪费功耗的一种方式，因此，尽量减少控制开销也可以提高功耗的效率。

（5）传输启动协议

直觉上，我们认为一个需要进行通信的终端首先必须发起传输。事实上，大多数协议就是按照这种方式进行的。不过，接收端启动协议更加适合某些特殊的网络，如传感器网络。在接收端启动协议中，接收端通过发送一个准备接收（Ready To Receive，RTR）数据包来轮询邻近的接收端，这样可以表示已经准备好进行数据接收。如果接收端可以知道或者成功预测相邻的接收端何时需要发送数据，那么接收端启动协议就可以实现很高的工作效率。但是，对于一般的网络和不可预见的流量来说，发送端启动协议是一个更好的选择。

(6) 流量负载和可升级性协议

在最坏的网络状况下，协议通常需要进行优化。稀疏的节点分布和低流量情况不会对 Ad Hoc 网络的实现带来任何挑战。对于高流量负载、高节点密度和实时流量网络来说，协议必须根据具体的应用来进行优化。能提供信道预留概率的协议，在高负载和实时流量网络中也能应对自如。接收端启动方法在高负载网络中也可以表现很好，因为 RTR 数据包到达需要发送数据的终端的概率非常高。如果网络对终端和流量进行分级，那么网络可以根据流量的性质来制定优先级，这样，就可以很好地处理实时流量了。由于传输节点优先级的原因，越密集的网络就越会产生干扰。基于这个原因，功耗控制的应用就会对网络的性能产生很大的影响。

(7) 传输范围

传输范围是指无线信号强度处于最低可用级别时，终端与发射天线之间的距离。协议可以分为超短距离（小于 10m）、短距离（10~100m）、中距离（100~1000m）以及长距离（大于 1000m）。在扩大传输范围和实现更高的空间容量之间，存在一个权衡问题，这一点在协议的设计过程中必须充分考虑到。

2. 工业标准协议

(1) IEEE 802.11

IEEE 802.11 系列标准（IEEE 1999a; IEEE 1999b; IEEE 1999c）可以看作是无线版本的以太网局域网（LAN）协议。IEEE 802.11a 标准工作在免许可的 5 GHz 频带上，可以提供高达 54Mbit/s 的数据传输速率。最经济的 IEEE 802.11b 标准工作在 2.4GHz 的工业、科学和医学（Industrial Scientific and Medical, ISM）频带上，可以提供高达 11Mbit/s 的数据传输速率。IEEE 802.11 工作组当前的工作方向是服务质量（Quality of Service, QoS）和安全（IEEE 802.11i）。IEEE 802.11 标准主要定义了 MAC 层和物理层（PHY）的规范。在不同的物理层上，现有的 IEEE 802.11 标准采用了相同的媒质接入机制。基本的（强制性）接入机制称为“分布式协同功能（Distributed Coordination Function, DCF）”。可选择的点协同功能（Point Coordination Function, PCF）也是一种接入机制，在该机制中，中心节点（接入点）会根据一个列表来轮询各个终端。DCF 既可用于平台式 Ad Hoc 拓扑结构，也可用于集中式拓扑结构；而 PCF 只能用于集中式拓扑结构。MAC 提供了两种类型的流量业务：强制异步数据业务和可选时间绑定业务。强制异步数据业务是基于尽力服务（Best Effort, BE）的，非常适合对延时不敏感的数据传输；而可选时间绑定业务是通过 PCF 来实现的。

DCF 采用了基于载波侦听多址接入（Carrier Sense Multiple Access, CSMA）技术的“先听后发（Listen-Before-Talk）”方案。一个需要发送数据包的终端首先必须侦听媒质的动态，如果侦测到信道处于空闲状态，终端就会等待一个

DCF 帧间空间（DCF Interframe Space，DIFS）时间间隔（在 IEEE 802.11a 中为 34μs）；如果信道在 DIFS 阶段仍然处于空闲状态，那么终端在 DIFS 一结束就可以开始发送数据包了。当发送端从目的地接收到一个确认（ACK）数据包后，这个传输过程就算成功完成了。但是，如果侦听到信道处于繁忙状态，协议就会执行冲突避免流程。在该流程中，在 DIFS 时间内侦听到信道为空闲状态之后，需要发送数据的终端就会等待一个额外的随机补偿时间；然后，如果在这个额外的时间内信道仍然处于空闲状态，那么该终端就可以开始发送数据了。补偿时间是多个间隙时间（在 IEEE 802.11a 中为 9 个）的组合，该时间是由每个基站单独决定的。任何新的传输过程都需要在 0 和竞争窗口（Contention Window，CW）之间选择一个随机数。补偿时间在当媒质处于竞争阶段时会逐渐减小，否则就会保持不变。因此，补偿时间在耗尽之前会经历数个媒质的繁忙周期。图 21-2 给出了一个补偿流程的例子。CW 的初始值为 CW_{min}（IEEE 802.11a 中为 15），由于所有终端都工作在 CW_{min} 值上，因此它们都具有相同的初始媒质接入优先级；也就是说，当发送的数据包没有得到确认时，任何传输都是失败的，这时发送端就会将其 CW 值加倍至最大值 CW_{max}（IEEE 802.11a 中为 1023），这样，当多个终端需要接入媒质时，这个较大的 CW 值就可以减小发生冲突的概率。为了解决隐蔽终端问题，IEEE 802.11 选择性地使用了 RTS/CTS 握手技术。RTS 和 CTS 数据包中包含了数据帧的时长以及对应的 ACK。接收到 RTS 或 CTS 的终端利用 ACK 信息来启动一个定时器（称为“网络分配矢量（Network Allocation Vector，NAV）”），该定时器负责提示媒质的繁忙时段。在连续的 RTS 帧和 CTS 帧之间，以及数据帧和 ACK 之间，各插入了一个短帧间空间（Short Interframe Space，SIFS）（IEEE 802.11 中为 16μs）。SIFS 比 DIFS 时间短，因此，终端在发送这些帧时具有接入媒质的优先权。

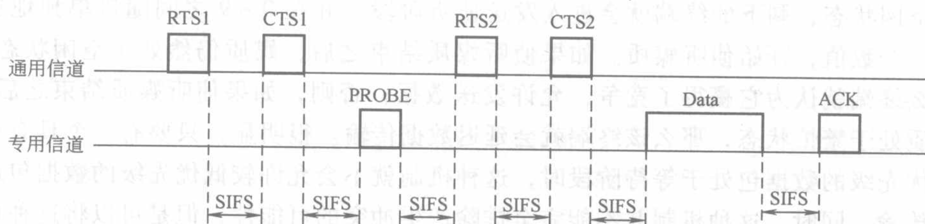


图 21-2 IEEE 802.11 补偿时间

（2）HIPERLAN 1

高性能 1 型 LAN（HIPERLAN 1）是一种无线 LAN 标准，该标准工作在 5 GHz 频带上，可以为簇式 Ad Hoc 或集中式拓扑结构的移动用户提供高达 23.5Mbit/s 的数据传输速率。HIPERLAN 1 可以提供具有分级优先权的异步尽力

服务和时间绑定业务，分级优先权总共包含5种优先级：0（最高）至4（最低）。每一个单独的MAC协议数据单元（Protocol Data Unit，PDU）都会被指定一个优先级，该优先级与PDU的标准剩余生命周期（Normalized Residual Lifetime，NRL）值非常接近。NRL是PDU根据在传输过程中还剩下多少跳数来对存活时间进行的估计。如果一个PDU的NRL值变为0，那么该PDU就会被丢弃。另外，有些终端作为数据转发器，负责将数据以多跳的方式中继至远节点。HIPERLAN 1允许终端进入休眠状态以节省电量，这些休眠的终端（称为"p-省电者"）负责向支持终端（称为"p-支持者"）告知其休眠/唤醒模式；然后p-支持者根据要求缓冲传向p-省电者的数据包。尽管HIPERLAN 1具有很多有用的特征，但是HIPERLAN 1在商业上并没有取得成功。HIPERLAN 1中使用的信道接入机制是非优先排除接入法（Elimination-Yield Nonpreemptive Priority Access，EY-NPA），该方法由3个阶段组成：区分优先级（在需要发送的数据包中确定具有最高优先级的数据包）、竞争（排除所有其他竞争者）、传输。在区分优先级阶段，时间被划分为5个迷你时隙，从0~4依次编号。需要发送数据的终端必须在迷你时隙内发送一个与其MAC PDU优先级对应的突发数据包。例如，具有两个优先PDU的终端会在迷你时隙0和1之间侦听媒质，如果媒质空闲，该终端会在迷你时隙2内通过发送突发数据包来提出数据传输请求。如果媒质在迷你时隙0或1之内变成了繁忙状态，那么该终端就会延迟数据传输。一旦发送了一个突发数据包，区分优先级阶段就宣告结束，剩下进行竞争的终端中的PDU都具有相同的优先级。接下来就是竞争阶段了。竞争阶段在开始时，竞争终端会发送一个排除突发数据包。各个独立的终端随机单独地选择突发数据包的长度，长度范围为0~12个迷你时隙。发送完突发数据包之后，终端就会侦听媒质；如果媒质处于繁忙状态，那么该终端就会延迟数据传输；否则，如果媒质处于空闲状态，剩下的终端就会进入发送侦听阶段，并在0~9之间随机单独地选择一个数值，开始侦听媒质。如果侦听媒质结束之后，媒质仍然处于空闲状态，那么终端就认为它赢得了竞争，允许发送数据；否则，如果侦听媒质结束之后，媒质处于繁忙状态，那么该终端就会延迟数据传输。很明显，只要有一个具有更高优先级的数据包处于等待阶段时，这种机制就不会允许较低优先级的数据包进行传输。同时，这种机制并不能完全排除产生冲突的可能性，但是可以将这种可能性降到最低。类似于IEEE 802.11，如果媒质处于空闲状态的时间比帧间时间长，那么需要发送数据的终端就会忽略EY-NPA法，而直接发送数据。

（3）蓝牙

蓝牙（Bluetooth，1999）是一种无线协议，工作在免许可ISM频段，用于连接移动设备和桌面设备，如计算机和计算机外围设备、手持设备、蜂窝电话等。蓝牙的目标是制造低成本、低功耗和超短距离通信设备，这些设备可以进行语音

和数据传输，最大速率为 1Mbit/s。蓝牙采用跳频扩频（Frequency Hopping Spread Spectrum，FHSS）技术，其频距为 1600 跳/s。在语音通信中，采用了一个速率为 64kbit/s 的全双工连接，称为“同步面向连接（Synchronous Connection Oriented，SCO）”，SCO 为点对点通信分配了一个周期单时隙；在数据通信中，采用了一个尽力服务异步无连接（Asynchronous Connectionless Link，ACL）方式，在该连接中，最多可以分配 5 个时隙。在蓝牙技术中，终端是以微微网（Piconet）的方式进行组织的。一个 Piconet 中包含一个主导终端和最多 7 个活动的附属终端。主导终端负责确定跳频模式，而其他终端必须与该主导终端保持同步。一个活动终端一旦进入一个 Piconet，该活动终端就会被分配一个惟一的 3bit 活动成员地址（Active Member Address，AMA）；之后，该活动终端就处于发送状态（参与通信时）或连接状态。蓝牙支持 3 种低功耗状态：停机、保留和寻找。处于停机状态的终端会释放它的 AMA，并分配一个 8bit 长的停机成员地址（Parked Member Address，PMA）；处于保留和寻找状态的终端会保留它们的 AMA，但被限制进入 Piconet。例如，一个处于保留状态的终端就不能利用 ACL 进行通信。一个不参与任何 Piconet 的终端处于待命状态。多个蓝牙 Piconet 可以在时间和空间上共存，因此一个终端可以属于多个 Piconet。当一个 Piconet 中的未来主导终端开始询问过程时，该 Piconet 就形成了；也就是说，该主导终端通过广播询问消息来发现附近的其他终端。在接收到询问响应之后，主导终端就会明确对加入 Piconet 的终端进行编号。如果主导终端已经知道其他终端的特征，该主导终端就会跳过询问阶段，直接对附属终端进行编号。蓝牙采用了时分复用（Time Divided Duplex，TDD）技术，在该技术中，主导终端和附属终端交替进行数据发送。如果主导终端刚刚传送过信号给一个附属终端，那么该附属终端就只能进行信号回传；也就是说，被主导终端刚刚轮询过的附属终端必须进行响应。尽管传输过程的典型时长为单时隙，但传输过程也可能会持续 1 个、3 个或 5 个时隙。

(4) IEEE 802.11e

IEEE 802.11e 是一种新兴的 MAC 协议，该协议在 IEEE 802.11 系列无线 LAN 标准的基础上定义了一组 QoS 特征。目前，该协议只有一个规范草案（IEEE，2003）。该协议的目标是为延时敏感的应用提供更好的服务，如语音和多媒体业务。在 IEEE 802.11e 中，基于竞争的媒质接入称为“增强分布式信道接入（Enhanced Distributed Channel Access，EDCA）”。为了适应不同的流量优先级，IEEE 802.11e 引入了 4 种接入类型（AC），每一种 AC 对应了一个补偿实体。每个 IEEE 802.11e 终端中的 4 种不同并行补偿实体分别为（从高至低的优先级）：语音、视频、尽力服务和背景。为了进行比较，现有的 IEEE 802.11/a/b 标准为每个终端只定义一个补偿实体，每个补偿实体都具有互不相同的一组参

数，如 CW_{min}、CW_{max} 和判优帧间空间（Arbitration InterFrame Space，AIFS）。AIFS 至少等于 DIFS，如果需要可以放大。另一个添加到 IEEE 802.11e 中的特征称为“传输机会（Transmission Opportunity，TxOP）”。一个 TxOP 定义了一个时间间隔，该间隔可以被补偿实体用来传输数据。TxOP 由开始时间和持续时间来描述，其最大长度与 AC 有关。IEEE 802.11e 协议还定义了每个 MAC 业务数据单元（MAC Service Data Unit，MSDU）的最大生命周期，这个生命周期同样与 AC 有关。一旦最大生命周期耗尽，MSDU 就会被丢弃。最后，协议包含了可选块确认机制，在该机制中，一串连续的 MSDU 是通过单个 ACK 帧来进行确认的。

3. 其他协议

PRMA——数据包预留多址接入协议（Goodman，1989）。在 PRMA 协议中，媒质被划分成各个时隙，每 N 个时隙形成一个帧，这些时隙既可以是预留的，也可以是被使用的。媒质的接入过程通过槽形 ALOHA 协议进行划分。数据可能是周期性的，也可能是不规则的，这一点在数据帧头中会有明确指示。当终端有周期性的数据需要发送时，终端可以预留一个时隙；一旦中心节点成功确认了周期性数据，终端就会认为该时隙是预留的，并无须竞争而直接使用该时隙。当终端停止发送周期性数据后，预留的时隙就会被释放。PRMA 协议假设了一个中心节点结构，但这种机制也适用于其他类型的拓扑结构（Jiang，2002）。

MACA-BI——基于邀请的冲突避免多址接入协议（Talucci，1997）。在 MACA-BI 协议中，接收端通过发送一个准备接收（RTR）数据包来轮询预期的发送端（这是接收端启动协议中的一个特例）。为了以时间的方式进行轮询，接收端必须正确预测发送端发起的流量。周期流量使这个任务变得更加简单了。当数据缓冲或者终端的延时增加到超过门限值时，该终端就会通过发送一个 RTS 数据包来触发一个通信过程。（Garcia，1999）提出了关于 MACA-BI 协议的改进措施，这些改进措施中介绍了接收端启动单轮询多址接入（Receive Initiated Multiple Access with Simple Polling，RIMA-SP）协议和接收端启动双重目的轮询多址接入（Receive Initiated Multiple Access with Dual-purpose Polling，RIMA-DP）协议，这两种协议都可以使 RTR 数据握手避免冲突。RIMA-DP 协议在 RTR 数据包的基础上添加了一个目标：来自轮询终端的传输请求。在预留阶段结束之后，终端之间会进行数据交换。

DBTMA——双忙音多址接入协议（Haas，2002）。在 DBTMA 协议中，RTS/CTS 握手被两个带外忙音代替了，即发送忙音（transmit Busy Tone，BTt）和接收忙音（receive Busy Tone，BTr）。当一个终端需要发送数据时，该终端首先必须侦听是否有 BTt 和 BTr 存在。如果媒质处于空闲状态（即没有检测到忙音），终端就会开启 BTt，并发送 RTS 数据包，同时关闭 BTr。在其他协议中，如果媒质处于繁忙状态，就会产生一个随机的补偿时间。目的终端在接收到发给该终端

的 RTS 之后，就会开启 BTt，并等待数据传输。一旦检测到了 BTt，发送端就会认为它已经成功占据了媒质。在等待一小段时间后（用于传输 BTr），发送端就会发送数据包。在成功接收到数据包后，目的终端就会关闭 BTr，结束通信过程；如果没有接收到数据，目的终端就会在一个定时器计时结束后关闭 BTr。

Fitzak 等人（Fitzak, 2003）在 IEEE 802.11 的基础上提出了一种多跳 MAC 协议。其中，公共信道负责传输信令，而专用信道负责传输数据流量和 ACK 数据包。图 21-3 给出了多跳协议的 MAC 握手过程。第一个 RTS 数据包用来联系目的终端，并表达接收数据的意愿。发送端拥有很多空闲的专用信道，目的终端从这些空闲的专用信道中选择一条，并将选择信息在一个 CTS 数据包中传输给发送端。如果没有合适的专用信道可用，那么握手过程就会结束。在接收到 CTS 数据包后，发送端会在专用信道上传输一个 PROBE 数据包；目的终端利用这个数据包来测试信道情况，并在公共信道上发送第二个 CTS 数据包，通知选用的编码/调制方案；发送端为了确认选所选的参数，会发送第二个 RTS 数据包。尽管这种方案更加复杂，但是方案的提出者声称这种方案比原始的 IEEE 802.11 效率更高。

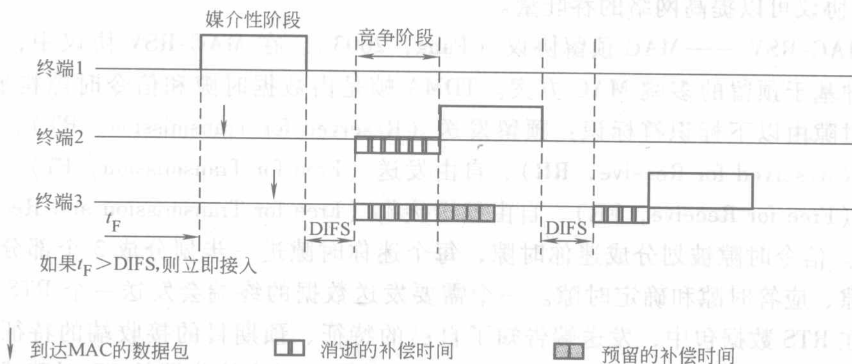


图 21-3 MAC 握手过程

LA-MAC——负载认知 MAC 协议（Chao, 2003）。在 LA-MAC 协议中，协议根据流量大小在竞争机制和免竞争机制之间转换。竞争机制最适合流量很小的情况，这种情况下，获取媒质时发生冲突的可能性非常小；对于流量很大的情况，免竞争机制可以获得更高和分布更均匀的吞吐量。在（Chao, 2003）中，IEEE 802.11 的 DCF 会自适应竞争阶段，而在免竞争阶段采用标记传送协议。流量负载是根据经历的延时数据包来衡量的，每个终端都会为将要发送的数据包计算这个延时。在竞争阶段，一个终端在发送数据包之前，会检查数据包的当前延时；如果这个延时大于预先定义的门限值 A，那么终端就会生成一个标记，并将该标记附着在数据包上一并发送。这就意味着，所有终端都会知道免竞争阶段的开始

点了。一旦这个延时降到低于一个预先定义的门限值 B，需要发送数据的终端就会去掉这个标记。这就意味着，在免竞争阶段结束时，竞争阶段就开始了。门限值 A 比门限值 B 大，这样可以描述一些滞后现象。

PCDC——功耗控制双信道协议（Muqattash，2003）。PCDC 协议的目标是以最低的传输功耗来维持网络的连接性。PCDC 是一个多跳协议，该协议使用了 IEEE 802.11 中提到过的 RTS/CTS 握手过程，并对其做了一些修改。每个终端都必须保存一列相邻的终端信息表以及可以到达相邻终端的传输功耗。当接收到一个数据包时，终端就需要访问这一列终端信息表了。如果在这一列终端信息表中找不到发送端，那么就必须添加一个新的终端信息；否则，就必须对现有终端信息表进行更新。任何情况下，接收端都必须对其连接信息进行重新评估，以确定该接收端知道到达相邻终端信息表中所有终端最划算的方式（发送功耗检测）。例如，对于一些终端来说，采用中间终端来代替直接路由就更加划算。在大流量负载的情况下，有足够的数据包来向所有终端告知关于它们相邻终端的信息。对于长时间处于空闲状态的终端来说，必须定期广播一个“HELLO”数据包来告知相邻的终端。PCDC 的空间效率和由 PCDC 协议提出者进行的仿真共同表明 PCDC 协议可以提高网络的吞吐量。

MAC-RSV ——MAC 预留协议（Fang，2003）。在 MAC-RSV 协议中，提出了一种基于预留的多跳 MAC 方案。TDMA 帧是由数据时隙和信令时隙构成的。数据时隙由以下标识符标识：预留发送（Reserved for Transmission，RT）、预留接收（Reserved for Receive，RR）、自由发送（Free for Transmission，FT）、自由接收（Free for Receive，FR）、自由发送接收（Free for Transmission and Receive，FTR）。信令时隙被划分成迷你时隙，每个迷你时隙进一步划分成 3 个部分：请求时隙、应答时隙和确定时隙。一个需要发送数据的终端会发送一个 RTS 数据包，在 RTS 数据包中，发送端告知了自己的特征、预期目的接收端的特征以及希望预留的数据时隙。如果任何被请求的时隙位于对应接收端的 FR 时隙或 FTR 时隙中，那么该接收端就会应答一个 CTS 数据包；否则，该接收端就会保持沉默。还存在一种可能，即 CTS 数据包只接受了一部分的时隙预留请求。如果一个请求时隙位于终端（除了预期目的终端）的 RR 时隙中，那么该终端就会应答一个 Not CTS（NCTS）数据包；检测到 RTS 冲突的终端同样通过发送一个 NCTS 来进行应答；否则，就保持沉默。最后，如果发送端成功接收到一个 CTS，那么该发送端就可以通过发送一个确定时隙（CONF）来确定预留；否则，该发送端就保持沉默。在请求迷你时隙时，需要发送 RTS 数据包；应答时需要使用 CTS 和 NCTS 数据包；确定时需要使用 CONF 数据包。数据时隙被划分为 3 个部分：接收端信标（Receiver Beacon，RB）、数据和确认信息（ACK）。数据时隙被标识为 RR 的终端会发送一个 RB，其中包含了活动数据发送端的特征。另外，接

收端在数据时隙结束时通过发送一个 ACK 来确认接收到正确的数据。由 MAC-RSV 协议提出者进行的仿真表明该协议在适应大流量负载方面比 IEEE 802.11 更优秀。

4. 注释

在（Jurdak，2004）中给出一组原则，即必须遵循合适的常用 MAC 协议。特别提到的是，必须使用多信道来隔离控制信号和数据信号，以便降低产生冲突的概率。灵活的信道带宽、多信道和高带宽效率的需求表明 CDMA 是信道分割的最佳选择。多跳协议可以确保应用中平台式拓扑或簇式拓扑结构的可升级性。为了支持高功效终端，协议必须对功耗非常清楚，而且必须能控制发送功耗，同时还能支持休眠模式。为了完善这些要求并满足灵活性，Jurdak 详细介绍了短距离网络和中距离网络以及发送端启动方法。

21.4 Ad Hoc 网络的 TCP

TCP 是目前互联网中非常流行的一种传输协议，Ad Hoc 网络中使用 TCP 也是一种必然趋势。这一点激发了人们对 TCP 的广泛研究，不仅加强了对 Ad Hoc 网络中 TCP 的性能进行评估，而且还为 Ad Hoc 网络提出了各种合适的 TCP 方案。TCP 最初是基于以下前提条件而为无线网络设计的：数据包丢失主要是由拥塞导致的、连接是稳定的（非常低的数据传输速率）、往返时间是稳定的、带宽是不变的（Postel，1981；Huston，2001）。基于上述的前提，TCP 流控制采用了窗口技术，窗口技术中的关键概念就是确定网络中的可利用资源。窗口根据相加性增加策略/相乘性减小策略来进行调整。当检测到数据包产生丢失时，TCP 发送端就会重新发送丢失的数据包，拥塞控制机制就会被调用，该机制中包含一个指数级的重发定时器补偿机制和一个通过缩小窗口大小来减小重发速率的机制。因此，数据包丢失作为产生拥塞的一种体现就可以通过 TCP 来进行描述（Chandran，2001）。前面关于蜂窝无线网络中 TCP 应用的研究表明，导致 TCP 效率低下主要是因为无线网络中产生数据包丢失的主要原因不再是拥塞现象了，而是容易产生错误的无线媒质（Xylomenos，2001；Balakrishnan，1997）。另外，无线网络中的多个用户可以共享相同的媒质，这样可以补偿传输延时时间差。因此，由于传输误差或延时数据包导致的数据包丢失可以通过 TCP 来描述，如同由于拥塞导致的数据包丢失一样。当 TCP 贯穿了整个 Ad Hoc 网络之后，就又会产生额外的问题。与蜂窝网络不同，Ad Hoc 网络中只有最后一跳是无线连接，而 TCP 发送端和接收端之间整个路径可能都是有线连接的（多跳）。因此，如本章之前所述，Ad Hoc 网络需要一个合适的路由协议和媒质接入控制机制（在链接控制层），来建立连接发送端和接收端的路径。TCP 和物理层协议、链接控制层和网

络层之间的相互作用可能会导致产生一些严重的性能减退，如下所述。

1. 物理层影响

干扰和传输信道效应是无线网络中产生高错误率的主要原因。由信道引起的错误可能会破坏TCP数据包或确认（ACK）数据包，从而造成数据包丢失。如果在重发超时（Retransmit TimeOut，RTO）间隔内没有收到ACK，那么数据包丢失就会被错误地认为是拥塞的表现，这种现象导致了TCP拥塞控制机制的产生。添加了TCP拥塞控制机制后，TCP发送速率就会大幅度下降，从而使得整体性能也会下降。因此，TCP对由于各种错误数据包丢失产生的反应动作很明显是不合适的。一种避免产生这种TCP行为的方法就是通过使用合适的前向纠错（Forward Error Correction，FEC）编码来使无线信道变得更加可靠，但代价是有效带宽会有所下降（由于额外的冗余）和传输延时有所增加（Shakkottai，2003）。除了FEC之外，链接控制层的自动重发请求（Automatic Repeat reQuest，ARQ）方案也可以用来提供比上面各层更快的重发速率。ARQ方案可能会增加传输延时，从而导致TCP假设了一个很大的往返时间，或者同时触发ARQ自身的重发流程（Huston，2001）。

2. MAC层影响

众所周知，隐藏和外露终端问题会严重降低Ad Hoc网络的整体性能。人们提出了各种技术来避免这个问题，包括IEEE 802.11 MAC协议中的RTS/CTS控制数据包交换。但是，尽管使用这些技术，隐藏和外露终端问题仍然还会出现，并导致反常的TCP行为。多跳情形下TCP和链接控制层机制之间不合适的相互作用会导致产生所谓的“TCP不稳定性”（Xu，2001）。TCP可以通过调整竞争窗口大小来自适应控制发送速率，窗口的大小决定了网络中正在传输的数据包的数量（例如，数据包的数量可以在接收到ACK之前通过TCP发送端发送）。较大的窗口会提高链接控制层中的竞争等级，因为更多的数据包需要寻找到达目的终端的路径。提高后的竞争等级会带来数据包冲突，并导致外露终端问题，从而阻碍了中间节点到达它们相邻的终端（Xu，2001）。当一个终端不能发送数据包给相邻终端时，该终端就会给源终端报告一个路由故障，这个过程是通过调用路由协议中的路由重建机制来实现的。如果路由重建耗费的时间比RTO长，那么TCP拥塞控制机制就会被触发，同时还会缩小窗口大小，并重发丢失的数据包。拥塞控制机制会导致TCP吞吐量产生瞬间下降，从而产生前面提到过的TCP不稳定性。实践证明，减小TCP竞争窗口的大小，可以降低TCP不稳定性（Xu，2001）。但是，减小后的窗口大小会抑制多跳情形下空间信道的重复使用性。对于IEEE 802.11 MAC（使用了4种握手方式（RTS、CTS、Data、ACK））协议，在一个H跳的链配置中，假设采用理想的时序安排和相同的窗口大小，最多只有$H/4$个终端可以同时发送数据（Fu，2003）。因此，小于上限的窗口会降低信

道的利用率。另一个与 TCP 和链接控制层机制之间相互作用相关的因素是多个 TCP 对话过程中的不公平问题。不公平问题（Xu，2001；Tang，2001）还与隐藏（冲突）和外露终端问题相关，不公平问题可以终止一个 TCP 对话。当一个终端由于冲突或外露终端问题而不能向相邻终端发送数据时，它在链接控制层中的补偿方案就会被激活，同时（随机）增加补偿时间。如果补偿方案反复被激活，终端就很难赢得竞争了；赢得竞争的终端将最终占据媒质，同时终止竞争失败终端的所有 TCP 对话。

3. 移动性影响

由于终端的移动，在 TCP 对话的生命周期内会时常产生路由故障。如前所述，当检测到一个路由故障后，路由协议会激活路由重建机制，如果发现新路由的耗时比 RTO 长，那么 TCP 发送端将会把路由故障认为成拥塞现象。这样，TCP 拥塞控制机制就会被激活，丢失的数据包就会被重发。但是，TCP 的这种反应行为很显然是不合适的，原因有如下几点（Chandran，2001）：首先，丢失的数据包在路由重建好之前不应该重发；其次，当路由最终恢复后，TCP 的慢启动策略将会迫使 TCP 的吞吐量在路由重建好之后，不必要地立即减小；另外，如果路由故障频繁发生，那么 TCP 的吞吐量将永远达不到最高速率值。

4. Ad Hoc 网络中的主要 TCP 方案

（1）TCP—反馈

这种 TCP 方案的基础是向 TCP 发送端明确告知路由故障机制，这样 TCP 发送端就不会错误地激活拥塞控制机制（Chandran，2001）。当一个中间终端检测到一个路由故障后，该终端将会向 TCP 发送端发送一个路由故障通知（Route Failure Notification，RFN），并记录这次事件。接收到 RFN 后，TCP 发送端将会过渡到一个"瞌睡"状态，并停止发送数据包、冻结 TCP 流控制窗口的大小并暂停定时器、开启一个路由故障定时器。当一个发送过 RFN 的中间终端发现一条新路由后，该终端将会向 TCP 发送端发送一个路由重建通知（Route Reestablishment Notification，RRN），TCP 发送端将脱离"瞌睡"状态，恢复正常操作。

（2）具有精确链接故障通知的 TCP

精确链接故障通知（Explict Link Failure Notification，ELFN）技术的原理是为 TCP 发送端提供链接或路由故障信息，从而阻止 TCP 对这样的故障做出反应，如拥塞现象（Holland，2002）。在这种方法中，ELFN 消息是由路由协议生成的，发送给 TCP 发送端的链接故障通知附着在 ELFN 消息上。当 TCP 发送端接收到链接故障通知后，该发送端将停止它的重发计时器，并定期检测网络来核对路由是否重建成功。当接收到一个 ACK 后，TCP 发送端将认为一条新的路由已经建立好，并恢复正常操作。

（3）Ad Hoc TCP

Ad Hoc TCP 的重要特征就是，标准 TCP 没有被修改，而是在 IP 层和 TCP 层之间插入了一个中间层，称为“Ad Hoc TCP 层（ATCP）”。因此，ATCP 对于 TCP 来说是无形的，而且安装了 ATCP 的终端与没有安装 ATCP 的终端之间可以协同工作。ATCP 是根据互联网控制消息协议（Internet Control Message Protocol，ICMP）和精确拥塞通知（Explicit Congestion Notification，ECN）机制提供的网络状态信息来工作的（Floyd，1994）。ECN 机制用来向 TCP 目的终端告知网络中的拥塞情况。ECN 位包含在 TCP 帧头中，该位开始时被 TCP 发送端设置为 0。当中间路由器检测到拥塞时，该 ECN 位就会被设置为 1。当 TCP 目的终端接收到一个 ECN 位设置为 1 的数据包后，该终端就会通知 TCP 发送端网络存在拥塞，这样，TCP 发送端就会降低发送速率。ATCP 具有 4 种状态：正常、拥塞、丢失和断开。在正常状态下，ATCP 不会有任何动作，对于 TCP 来说是隐形的；在拥塞、丢失和断开状态下，ATCP 将会分别处理拥塞的网络、丢失的信道和被断开的网络。当 ATCP 遇到 3 个复杂的 ACK（由信道引发的错误导致的）时，ATCP 将会过渡到丢失状态，并将 TCP 放入持续模式，以确保 TCP 不会激活拥塞控制机制。在丢失状态下，ATCP 将会重发未被确认的数据段。当接收到一个新的 ACK 后，ATCP 将会返回到正常状态，并将 TCP 从持续模式中释放出来，同时恢复 TCP 的正常操作。当网络发生拥塞时，ATCP 将会根据设置为 1 的 ECN 位来过渡到拥塞状态；在拥塞状态下，ATCP 不会影响 TCP 的拥塞控制机制。最终，当出现路由故障时，ICMP 将会发送一个未到达目的终端的消息。在持续模式中，TCP 将会定期发送一个检测数据包。当路由重建好之后，TCP 将会从持续模式中释放出来，ATCP 将会恢复到正常状态。

21.5 Ad Hoc 网络的容量

由香农（Shannon，1948）创立的经典信息论给出了信道容量的理论结果，即在狭窄的噪声通信信道中能传输多少信息。在 Ad Hoc 网络中，这个问题上升到了一个更高的难度，因为容量必须根据发送端和接收端来进行测量。无线网络的容量分析具有与经典信息论相似的目标，即评估信息发送的最高速率并确定最佳操作模式，以便达到最高发送速率。这个最高速率首先是由 Gupta 和 Kumar 在（Gupta，2000）中计算出来的；在这个计算过程中，两位作者在以下假设条件的前提下，为静态 Ad Hoc（节点不移动）网络容量的研究提出了一种模型。假设各个节点位于区域 1 中，每个节点可以在公共无线信道上以一定的数据传输速率发送数据，数据包在节点之间以多跳的方式发送，直至目的节点；而且数据包也可以在中间节点进行缓冲，以等待发送。作者还假设了两种网络配置：任意网络（其中，节点位置、流量目的地、速率和功耗等级都是任意的）和随机网络

（其中，节点位置和流量目的地是随机的）；但这两种网络具有相同的发送功率和数据传输速率。下面介绍了在单跳情形下的两个成功接收模型：

1）协议模型：在该模型中，如果 $d_{kj} \geqslant (1+\Delta) d_{ij}$，那么节点 i 至节点 j 的传输（两个节点之间的距离为 d_{ij}）就是成功的；也就是说，当节点 i 和节点 k 在相同的信道上同时向节点 j 发送数据时，如果节点 i 与节点 j 之间的距离小于节点 k 与节点 j 之间的距离，那么节点 i 至节点 j 的传输就是成功的。变量 $\Delta > 0$ 模拟了协议指定的监视区域，用来阻止相邻的节点同时发送数据。

2）物理模型：在该模型中，T 是所有同时发送的节点的一个子集，如果满足下面的条件，就表明节点 j 成功接收到了节点 i 发送的数据（$i \in T$）：

$$\frac{P_i / d_{ij}^{\alpha}}{N + \sum_{\substack{k \in T \\ k \neq i}} P_k / d_{kj}^{\alpha}} \geqslant \beta \tag{21-1}$$

式中，β 为成功接收时的最小信干比（Signal-to-Interference Ratio，SIR）；N 为噪声功率等级；P_i 为节点 i 的发送功率等级；α 为信号功率衰减。

传输能力被定义为一定距离下传输的数据数量，单位为 bit · m。1bit · m 表示 1bit 被传输到 1m 远的目的地。在前面介绍的接收模式中，任意网络中的传输能力上限和随机网络中每个节点的吞吐量就可以计算出来了，如表 21-1 所示；在该表中，使用了 Knuth 符号，即 $f(n) = \Theta(g(n))$ 表示 $f(n) = O(g(n))$，$g(n) = O(f(n))$；表中，c 和 c' 为常量，该常量是 α 和 β 的函数。表中的这些结果表明，对于任意网络来说，如果传输能力在所有节点中被划分成各个相等的部分，那么每个节点的吞吐量将是 $\Theta(W/\sqrt{n})$ bit/s；这就意味着，随着节点数量的增加，每个节点的吞吐能力将会以平方根比例减小。对于随机网络，结果是相同的。这些结果假设了一个完美的规划原则，该原则知道所有节点的位置和所有的流量需求，并在时间和空间上对无线传输进行了调整，以避免冲突。没有这些假设条件，上面的传输能力值将会变小。

表 21-1　任意网络和随机网络的容量上限

	协议模型	物理模型
任意网络（传输能力）/（bit · m/s）	$\Theta(W/\sqrt{n})$	在适当的 c' 和 c 值条件下，$cW/\sqrt{n}$ 切实可行；$c'Wn^{\left(\frac{\alpha-1}{\alpha}\right)}$ 不可行在适当的 c' 和 c 值条件下
随机网络（节点吞吐量）/（bit/s）	$\Theta\left(\frac{W}{\sqrt{n\lg n}}\right)$	$cW/\sqrt{n\lg n}$ 切实可行；$c'W/\sqrt{n}$ 不可行

Troumpis 和 Goldsmith（Troumpis，2000）将传输能力上限分析扩展到了三维拓扑结构中，并将信道容量融合到了链接模型中了。在这个过程中，各个节点假设是均匀分布在一个 $1\mathrm{m}^3$ 的立方区域中。容量 $C(n)$ 满足下面的不等式：

$$k_1\frac{n^{1/3}}{\lg(n)}\leqslant C(n)\leqslant k_2\lg(n)n^{1/2} \tag{21-2}$$

当 $n\to\infty$ 时，容量趋于不变；其中，k_1 和 k_2 为正数常量。式（21-2）还说明，尽管容量会随着用户数量的增加而增加，但每个用户的可用速率将会下降。

Ad Hoc 网络容量的研究示例

（1）IEEE 802.11

Li 等人（Li，2001）通过仿真和区域测试研究了 Ad Hoc 网络的容量。此外，静态 Ad Hoc 网络是基本条件，这一点可以通过以下事实得到证实，即在大多数移动环境中，各个节点在数据传输时不会移动很大的距离。IEEE 802.11MAC 协议用来分析不同配置网络的容量。

对于一连串的节点来说，理想的容量为自然信道带宽的 1/4，该自然信道带宽可以通过无线链路实现（单跳吞吐量）。基于 IEEE 802.11 的仿真 Ad Hoc 网络的容量可以达到单跳吞吐量的 1/7，因为 IEEE 802.11 协议无法发现传输的最佳方案，而且其补偿流程在 Ad Hoc 网络的促进下执行起来仍然非常困难。区域训练得到的结果与仿真过程中得到的结果相同，这些相同的结果是在格状拓扑结构中发现的。对于具有随机流量模式的随机网络来说，IEEE 802.11 协议的效率已经不够了，但是每个节点在理论上可以达到 $O(1/\sqrt{n})$ 的最高容量。

另外，Ad Hoc 网络的可升级性是流量模式的一个函数。为了使整体容量随着网络规模的增加而增加，当网络规模逐渐增大时，源节点和目的节点之间的平均距离必须保持很小。因此，决定大型网络是否切实可行的关键因素就是流量模式。对于局部化流量的网络来说，网络扩展是切实可行的，但是对于流量必须完全贯穿整个网络的网络来说，扩展的可行性就有些值得疑问了。

（2）无线 Mesh 网络

作为 Ad Hoc 网络的一种特殊情况，无线 Mesh 网络（Wireless Mesh Network，WMN）吸引了很多人的关注。WMN 与其他 Ad Hoc 网络的主要区别是流量模式：在实际中，所有的流量都是来自或流向一个节点（网关），该节点连接到其他网络（如互联网）。因此，网关在 WMN 中扮演了一个决定性的角色：网关数量越多，网络的容量越大，网络的可靠性越高。Jun 和 Sichitiu（Jun，（Oct）2003）利用固定节点分析了 WMN 的容量，他们的分析表明 WMN 的容量主要取决于以下几个方面：

1）转发流量和顺畅性——WMN中的每个节点必须发送转发流量，如同自身流量一样。因此，在节点自身流量和转发流量之间必然存在一种竞争关系。实际上，随着每个节点提供的负载的增加，各个接近网关的节点就可以占用更多的带宽，即使在公平的MAC层协议中。绝对的公平性必须根据提供的负载来强制执行。

2）MAC层的理论容量（B）——定义为单跳网络MAC层中可达到的最大吞吐量，其计算值可以参考（Jun（April）2003）。

3）链接约束条件和冲突域——大体上，所有的MAC协议都被设计成可以避免冲突，同时确保只有一个节点在一个时刻和给定区域中发送数据。冲突域是指一组链接（包括正在发送数据的链接），这些链接必须是不活动的，以便某个链接发送成功。

首先分析的是链形拓扑。我们观察到，接近网关的节点必须比那些远离网关的节点多发送一些流量。对于包含n个节点的网络（其中每个节点产生负载流量为G）来说，网关与相邻节点之间的链接必须能够发送nG的流量；一个节点与下一个节点之间的链接必须能够发送$(n-1)G$的流量，依此类推。这样，冲突域可以区别出来，而且瓶颈冲突域（传输网络中的大部分流量）也可以确定下来。每个节点的有效吞吐量受到理论容量B的限制，该理论容量被瓶颈冲突域的总流量划分。链形拓扑分析可以拓展到二维拓扑结构（任意网络）中去，该分析中得到的每个节点的吞吐量得到了仿真结果的确认。

这些分析结果给出了每个节点吞吐量的渐近值$O(1/n)$，这个结果远比表21-1中的结果恶劣，这主要是因为网关的存在，网关就是网络的瓶颈。很明显，有效吞吐量会随着网络中网关数量的增加而提高。

（3）提高Ad Hoc网络的容量

表21-1中的描述说明了在静态网络中假设最佳的时序安排、路由和数据转发，就可以得到最佳的性能。对于网络规模拓展来说，这是一种不好的想法，除非网络的可升级性得到充分关注，并鼓励研究人员开发可以提高平均吞吐量的新技术。一种提高网络容量的方法就是在网络中添加转发节点（只有转发功能），这种方案的主要缺陷是它需要大量的纯转发节点。对于协议模型中包含m个额外转发节点的随机网络来说，每个节点的有效吞吐量就变成了$\Theta(W(n+m)/n\sqrt{(n+m)\lg(n+m)})$（Gupta，2000）。例如，在一个包含100个发送端的网络中，至少需要4476个转发节点来实现5倍的容量（Gupta，2000）。另一种策略就是在模型中引入移动性。Grossglauser和Tse（Grossglauser，2001）的研究表明每个发送-接收对都可以得到总有效带宽中的一部分固定带宽，该带宽与发送-接收对的数量无关，但代价是数据包的传输会产生延时，每个中间转发节点的缓冲区域也会增加。Bansal和Liu（Bansal，2003）也得到过相同的结果，

但是其延时约束条件很低而且移动性模型类似于随机中转模型（Bettstetter，2002）。移动性又带来了新的问题，例如在Ad Hoc网络中保持连接性、分发路由信息、建立接入控制。各个节点也可以划分成各个簇，在每个簇中，都有一个指定的节点（簇头）来承载转发数据（Lin，1997）。这种方法可以提高节点的容量，并可以降低由于路由和MAC协议产生的传输开销带来的影响。另一方面，簇头中的信息更新机制也会产生额外的传输开销，这种开销会降低节点的有效吞吐量。

参考文献

[1] Aggelou, G., Tafazolli, R., RDMAR: A Bandwidth-efficient Routing Protocol for Mobile Ad Hoc Networks, *ACM International Workshop on Wireless Mobile Multimedia* (*WoWMoM*), 1999, pp. 26-33.

[2] Balakrishnan, H., Padmanabhan, V. N., Seshan, S., Katz, R. H. A, A Comparison of Mechanisms for Improving TCP Performance over Wireless Links, IEEE/ACM Trans. On Networking, vol. 5, No. 6, pp. 756-769, December 1997.

[3] Bansal, N. and Liu, Z. Capacity, Delay and Mobility in Wireless Ad-Hoc Networks, *Proc. IEEE Infocom*'03, April 2003.

[4] Basagni, S., et al., A Distance Routing Effect Algorithm for Mobility (DREAM), *ACM/IEEE Int' l. Conf Mobile Comp. Net.*, pp. 76-84, 1998.

[5] Bellur, B. and Ogier, R. G., A Reliable, Efficient Topology Broadcast Protocol for Dynamic Networks, *Proc. IEEE INFOCOM*' 99, New York, March, 1999.

[6] Bettstetter, C., On the Minimum Node Degree and Connectivity of a Wireless Multihop Network. *Proc. ACM Intern. Symp. On Mobile Ad Hoc Networking and Computing* (*MobiHoc*), June 2002.

[7] Bluetooth SIG, Specification of the Bluetooth System, vol. 1. 0, 1999, available at: http://www.bluetooth.org.

[8] Chandra, A., Gummalla, V., Limb, J. O., Wireless Medium Access Control Protocols, *IEEE Communications Surveys and Tutorials* [*online*], vol. 3, no. 2, 2000, available at: http://www.comsoc.org/pubs/surveys/.

[9] Chandra, K., Raghunathan, S., Venkaresan, S., Prakash, R., A Feedback-Based Scheme for Improving TCP Performance in Ad Hoc Wireless Networks, *IEEE Personal Communications*, pp. 34-39, February 2001.

[10] Chao, C. M., Sheu, J. P., Chou, I-C., A load awareness medium access control protocol for wireless Ad Hoc network, *IEEE International Conference on Communications*, ICC' 03, vol. 1, 11-15, pp. 438-442, May 2003.

[11] Chatschik, B., A overview of the Bluetooth wireless technology, *IEEE Communications Magazine*, vol. 39, no. 12, pp. 86-94, Dec. 2001.

[12] Chen, T. -W., Gerla, M., Global State Routing: a New Routing Scheme for Ad-Hoc Wireless Networks, *Proceedings of the IEEE ICC*, 1998.

[13] Chiang, C. -C. and Gerla, M., Routing and Multicast in Multihop, Mobile Wireless Networks, *Proc. IEEE ICUPC*' 97, San Diego, CA, Oct. 1997.

[14] Corson, M. S. and Ephremides, A., A Distributed Routing Algorithm for Mobile Wireless Networks,

ACM/Baltzer Wireless Networks 1, vol 1, pp. 61-81, 1995.

[15] Das, S., Perkins, C., and Royer, E., Ad Hoc on Demand Distance Vector (AODV) Routing, Internet Draft, draft-ietf-manet-aodv-11. txt, 2002.

[16] Dube, R., Rais, C., Wang, K., and Tripathi, S., Signal Stability Based Adaptive Routing (SSA) for Ad Hoc Mobile Networks, *IEEE Personal Communication* 4, vol 1, pp. 36-45, 1997.

[17] Fang, J. C. and Kondylis G. D., A synchronous, reservation based medium access control protocol for multihop wireless networks, 2003 *IEEE Wireless Communications and Networking*, *WCNC* 2003, vol. 2, 16-20. March 2003, pp. 994-998.

[18] Fitzek, F. H. P., Angelini, D., Mazzini, G., and Zorzi, M., Design and performance of an enhanced IEEE 802. 11 MAC protocol for multihop coverage extension, *IEEE Wireless Communications*, vol. 10, no. 6, pp. 30-39, Dec. 2003.

[19] Floyd, S., TCP and Explicit Congestion Notification, *ACM Computer Communication Review*, vol. 24, pp. 10-23, October 1994.

[20] Fu, Z., Zerfos, P., Luo, H., Lu, S., Zhang, L., and Gerla, M., The Impact of Multihop Wireless Channel on TCP Throughput and Loss, *IEEE INFOCOM*, pp. 1744-1753, 2003.

[21] Garcia-Luna-Aceves, J. J. and Marcelo Spohn, C., Souce-tree Routing in Wireless Networks, *Proceedings of the Seventh Annual International Conference on Networks Protocols*. Toronto, Canada, p. 273, October 1999.

[22] Garcia-Luna-Aceves, J. J. and Tzamaloukas, A., Reversing the Collision-Avoidance Handshake in Wireless Networks, *ACM/IEEE MobiCom' 99*, pp. 15-20, August 1999.

[23] Goodman, D. J., Valenzuela, R. A., Gayliard, K. T., Ramamurthi, B., Packet reservation multiple access for local wireless communications, *IEEE Transactions on Communications*, vol. 37, no. 8, pp. 885-890, Aug. 1989.

[24] Grossglauser, M. and Tse, D., Mobility Increase the Capacity of Ad Hoc Wireless Networks, *Proc. IEEE Infocom' 01*, April 2001.

[25] Günes, M., Sorges, U., and Bouazizi, I., ARA-The Ant-colony Based Routing Algorithm for Manets, *ICPP Workshop on Ad Hoc Networks* (*IWAHN* 2002), pp. 79-85, August 2002.

[26] Gupta, P. and Kumar, P. R., The Capacity of Wireless Networks, *IEEE Trans. Info. Theory*, vol. 46, March 2000.

[27] Haas, Z. J. and Deng, J., Dual busy tone multiple access (DBTMA)—a multiple access control scheme for Ad Hoc networks, *IEEE Transactions on Communications*, vol. 50, no. 6, pp. 975-985, June 2002.

[28] Haas, Z. J. and Pearlman, R., Zone Routing Protocol for Ad-hoc Networks, *Internet Draft*, draft-ietf-manet-zrp-02. txt, 1999.

[29] Holland, G., and Vaidya, N., Analysis of TCP Performance over Mobile Ad Hoc Networks, *Wireless Networks*, Kluwer Academic Publishers, vol. 8, pp. 275-288, 2002.

[30] Huston, G., TCP in a Wireless World, *IEEE Internet Computing*, pp. 82-84. March-April, 2001.

[31] IEEE 802. 11 WG, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, *IEEE/ANSI Std*. 802-11, 1999 edn.

[32] IEEE 802. 11 WG, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: high-speed physical layer in the 5 GHz band, *IEEE Std*. 802-11a.

[33] IEEE 802. 11 WG, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: high-speed physical layer extension in the 2. 4 GHz band, *IEEE Std*. 802-11b.

[34] IEEE 802. 11 WG, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS), *IEEE Draft Std*. 802. 11e/D5. 0, August 2003.

[35] Iwata, A. et al., Scalable Routing Strategies for Ad-hoc Wireless Net-works, *IEEE JSAC*, pp. 1369-1379, August. 1999.

[36] Jacquet, P., Muhlethaler, P., Clausen, T., Laouiti, A., Qayyum, A., and Viennot, L., Optimized Link State Routing Protocol for Ad Hoc Networks, *IEEE INMIC*, Pakistan, 2001.

[37] Jiang, M., Ji, J., and Tay, Y. C. Cluster based routing protocol, *Internet Draft*, draft-ietf-manet-cbrp-spec-01. txt, 1999.

[38] Jiang, S., Rao, J., He, D., Ling, X., and Ko, C. C., A simple distributed PRMA for MANETs, IEEE Transactions on Vehicular Technology, vol. 51, no. 2, pp. 293-305, March 2002.

[39] Joa-Ng, M. and Lu, I. -T., A Peer-to-peer Zone-based Two-level Link State Routing for Mobile Ad Hoc Networks, *IEEE Journal on Selected Areas in Communications* 17, vol. 8, pp. 1415-1425, 1999.

[40] Johnson, D., Maltz, D., and Jetcheva, J., The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks, *Internet Draft*, draft-ietf-manet-dsr-07. txt, 2002.

[41] Jun, J. and Sichitiu, M. L., The Nominal Capacity of Wireless Mesh Networks, *IEEE Wireless Communications*, October 2003.

[42] Jun, J., Peddabachagari, P., and Sichitiu, M. L., Theoretical Maximum Throughput of IEEE 802. 11 and its Applications, Proc. *2nd IEEE Int'l Symp. Net. Comp. and Applications*, pp. 212-215, April 2003.

[43] Jurdak, R., Lopes, C. V., and Baldi, P., A Survey, Classification and Comparative Analysis of Medium Access Control Protocols for Ad Hoc Networks, IEEE Communications Surveys and Tutorials [online], vol. 6, no. 1, 2004. available at: http: //www. comsoc. org/pubs/surveys/.

[44] Kasera, K. K. and Ramanathan, R., A Location Management Protocol for Hierarchically Organized Multihop Mobile Wireless Networks, *Proceedings of the IEEE ICUPC'* 97, San Diego, CA, pp. 158-162, October 1997.

[45] Ko, Y. -B., Vaidya, N. H., Location-aided Routing (LAR) in Mobile Ad Hoc Networks, *Proceedings of the Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking* (*Mobicom'* 98), Dallas, TX, 1998.

[46] Li, J., Blake, C., De Couto, D. S. J., Lee, H. I., and Morris, R., Capacity of Ad Hoc Wireless Networks, *Proc. 7th ACM Int'l Conf. Mobile Comp. and Net.*, pp. 61-69, July 2001.

[47] Lin, C. R. and Gerla, M., Adaptative Clustering for Mobile Wireless Networks, *IEEE Journal on Selected Ares in Communications*, vol. 15, pp. 1265-1275, September 1997.

[48] Mangold, S., Sunghyun Choi, Hiertz, G. R., Klein, O., and Walke, B., Analysis of IEEE 802. 11e for QoS support in wireless LANs, *IEEE Wireless Communications*, vol. 10, no. 6, pp. 40-50, December 2003.

[49] Muqattash, A. and Krunz, M., Power controlled dual channel (PCDC) medium access protocol for wireless Ad Hoc networks, *22nd. Annual Joint Conference of the IEEE Computer and Communications Societies*, *INFOCOM* 2003, vol. 1, pp. 470-480, 30 March-3 April 2003.

[50] Murthy, S. and Garcia-Lunas-Aceves, J. J., An Efficient Routing Protocol for Wireless Networks, *ACM Mobile Networks and App. J.*, Special Issue on Routing in Mobile Communication Networks, pp. 183-197, Oct. 1996.

[51] Nikaein, N., Laboid, H., and Bonnet, C., Distributed Dynamic Routing Algorithm (DDR) for Mobile Ad Hoc Networks, *Proceedings of the MobiHOC* 2000: *First Annual Workshop on Mobile Ad Hoc Networking and Computing*, 2000.

[52] Ogier, R. G. et al., Topology Broadcast based on Reserve-Path Forwarding (TBRPF), draft-ietf-manet-tbrpf-05. txt, INTERNET-DRAFT, MANET Working Group, March 2002.

[53] Park, V. D. and Corson, M. S., A Highly Adaptive Distributed Routing Algorithm for Mobile Wireless Networks, *Proceedings of INFOCOM*, April 1997.

[54] Pei, G., Gerla, M., and Chen, T. -W., Fisheye State Routing: A Routing Scheme for Ad Hoc Wireless Networks, *Proc. ICC* 2000, New Orleans, LA, June 2000.

[55] Pei, G., Gerla, M., Hong, X., and Chiang, C., A Wireless Hierarchical Routing Protocol with Group Mobile, *Proceedings of Wireless Communications and Networking*, New Orleans, 1999.

[56] Perkins, C. E. and Bhagwat, P., Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers, *Comp. Commun. Rev.*, pp. 234-244, Oct. 1994.

[57] Postel, J., Transmission Control Protocol, *IETF RFC* 793, September 1981.

[58] Radhakrishnan, S., Rao, N. S. V., Racherla, G., Sekharan, C. N. and Batsell, S. G., DST-A Routing Protocol for Ad Hoc Networks Using Distributed Spanning Trees, *IEEE Wireless Communications and Networking Conference*, New Orleans, 1999.

[59] Raju, J. and Garcia-Luna-Aceves, J., A New Approach to On-demand Loop-free Multipath Routing, *Proceedings of the 8th Annual IEEE International Conference on Computer Communications and Networks* (*IC-CCN*), pp. 522-527, Boston, MA, October 1999.

[60] Santivanez, C., Ramanathan, R., and Stavrakakis, I., Making Link-State Routing Scale for Ad Hoc Networks, *Proc.* 2001 *ACM Int'l. Symp. Mobile Ad Hoc Net. Comp.*, Long Beach, CA, October 2001.

[61] Schiller, J., Mobile Communications, Addison-Wesley, Reading, MA, 2000.

[62] Shakkottai, S., Rappaport, T. S. and Karlsson, P. C, Cross-Layer Design for Wireless Networks, *IEEE Communications Magazine*, pp. 74-80, October 2003.

[63] Shannon, C. E., A mathematical theory of communication, *Bell System Technical Journal*, vol. 79, pp 379-423, July 1948.

[64] Su, W. and Gerla, M., IPv6 Flow Handoff in Ad-hoc Wireless Networks Using Mobility Prediction, *IEEE Global Communications Conference*, Rio de Janeiro, Brazil, pp. 271-275, December 1999.

[65] Talucci, F., Gerla, M., and Fratta, L., MACA-BI (MACA By Invitation)-a receiver oriented access protocol for wireless multihop networks, *8th IEEE International Symposium on Personal*, *Indoor and Mobile Radio Communications*, PIMRC' 97, Vol. 2, pp. 435-439, 1-4 Sept. 1997.

[66] Tang, K., Gerla, M., and Correa, M., Effects of Ad Hoc MAC Layer Medium Access Mechanisms under TCP, *Mobile Networks and Applications*, Kluwer Academic Publishers, Vol. 6, pp. 317-329, 2001.

[67] Toh, C., A Novel Distributed Routing Protocol to Support Ad-hoc Mobile Computing, *IEEE 15th Annual International Phoenix Conf.*, pp 480-486, 1996.

[68] Toumpis, S. and Goldsmith, A., Ad Hoc Network Capacity, Conference Record of Thirty-fourth Asilo-

mar Conference on Signal Systems and Computers, Vol. 2, pp. 1265-1269, 2000.

[69] Tseng, Y. C. and Hsieh, T. Y., Fully power-aware and location-aware protocols for wireless multi-hop Ad Hoc networks, 11*th*. *International Conference on Computer Communications and Networks*, pp. 608-613, 14-16 Oct. 2002.

[70] Woo, S. -C. and Singh, S., Scalable Routing Protocol for Ad Hoc Networks, *Wireless Networks* 7, Vol. 5, 513-529, 2001.

[71] Xu, S. and Saadawi, T., Does the IEEE 802. 11 MAC Protocol Work Well in Multihop Wireless Ad Hoc Networks? *IEEE Communications Magazine*, pp. 130-137, June 2001.

[72] Xylomenos, G., Polyzos, G. C., Mahonen, P., and Saaranen, M., TCP Performance Issues over Wireless Links, *IEEE Communications Magazine*, Vol. 39, No. 4, pp. 53-58, April 2001.

第22章 网络通信

James E. Goldman

22.1 网络分析与设计的一般原则

1. 自顶向下面向业务方式的应用

网络通信是指通过合理利用硬件、软件和媒体来实现数据、语音、视频、图像或传真的传输。在业务应用中，网络通信是指在正确的时间和地点，以正确的成本将正确的信息传递到正确的决策者手中。由于在网络的分析、设计和实现过程中存在太多的变数，因此网络必须使用结构化的方法，以确保实现的网络能满足预定业务、组织或个人的通信需求。

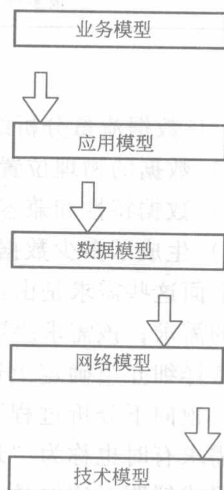


图 22-1 自顶向下设计方法：业务数据通信分析

结构化的方法称为“自顶向下法”，这种方法可以通过一个自顶向下的模型来说明，如图 22-1 所示。在自顶向下模型中采用自顶向下的方法相对直接一些。

首先，我们必须从业务级目标开始。一个公司（组织、个人）在构建网络时要实现的目标是什么呢？如果没有清楚掌握什么是业务级目标，就几乎不可能配置和构建一个成功的网络。

在完全理解了什么是业务级目标后，我们还必须掌握网络中计算机系统上运行的各种应用系统。因为，正是应用系统生成了网络中传输的各种流量。

在掌握了各种应用系统并记录之后，就要对这些应用系统生成的数据进行测试。在这种情况下，数据就是指目前网络中传输的各种有效负载，如语音、视频、图像、传真以及纯数据。数据流量分析不仅要确定等待传输的数据数量，而且要确定这些数据的重要本质特性。图 22-2 给出了数据流量分析的概要。在自顶向下的分析过程中，还必须检查网络中各节点的地理相邻性。

地理相邻性是各种不同类型网络的一个区别性因素，随后将对其及进行进一步的检查。

数据流量分析类型	数据流量分析描述
负载类型分析	大多数应用至少需要语音和数据业务，也有可能需要视频会议和多媒体，在选择电路和网络硬件之前，必须考虑到所有的负载类型，并记录下来
业务分析	利用过程流分析或文件流分析来区分各种业务类型，并分析每种业务类型的数据需求
时间分析	一旦所有业务类型区分完毕之后，就分析每个业务类型什么时候、多频繁地出现
流量大小分析	通过结合时间分析的结果来分析所有类型的业务，就可以得到一个时间敏感的流量大小需求规范，这将是电路容量与带宽进行映射的起点
关键任务分析	该分析的结果将明确说明专用数据安全流程的需求，如加密法或专用网络设计注意事项（冗余链接）
协议栈分析	分析每个企业位置的数据流量以及协议，该协议必须贯穿企业的广域网；存在很多各种各样的可选传输协议，但首先必须区分清楚这些协议

图 22-2　数据流量分析

一旦数据流量分析结束，我们必须知道以下几个问题的答案：

1）数据的物理位置在哪里？

2）数据特性和兼容性要点是什么？

3）生成了多少数据，有多少数据等待传输？

上面这些需求是由自顶向下模型中的各个上层决定的，下一项工作就是确定网络的需求；该需求决定了网络及时、有效划算地传递数据的能力。本节剩下的部分将详细介绍确定上述需求的具体细节。网络性能标准可以描述成“为了满足在自顶向下分析过程开始阶段描述的业务目标需求，网络必须完成的任务”。这些需求有时也称为“逻辑网络设计”。

形成鲜明对比的是，技术层分析将决定各种硬件和软件如何进行组合来构建一个功能性的网络，该网络可以满足预定的业务目标。所需技术的具体描述过程通常称为“物理网络设计”。

总之，自顶向下模型中各层之间的关系可以描述如下：上层分析可以提出各种需求，这些需求从上向下传递，而满足这些需求的解决方法将从下向上反馈到上层。如果各层之间的这种关系在整个面向业务的网络分析过程中都非常有效的话，那么底层实现技术应该能实现顶层开始提出的业务目标。

2. 开放式系统互联模型的应用

为了决定使用哪一种技术来满足逻辑网络设计（网络层）中确定的各种需求，我们需要一个结构化的方法。幸运的是，国际标准化组织（International Organization for Standardization，ISO）开发出了一个用于组织网络技术方案的基本

结构，称为“开放式系统互联（Open System Interconnect，OSI）参考模型”，如图 22-3 所示。

各层	主要功能	形象的比喻说明
7 应用层	包含了网络操作系统和应用程序，也是用户接口层	家具装饰品：椅子、床、桌子、装饰画
6 表达层	确保各个应用之间对话传输的可靠性；注意各种应用之间的区别；负责数据表达	内部木工作业：橱柜、架子、模塑
5 对话层	使能两个应用程序在整个网络中进行通信	电气应用：光学开关和光学线路之间的连接
4 传输层	确保端到端的可靠传输，通常会经过多个节点	加热/制冷/管道工程：熔炉、A/C 转换器、管道系统
3 网络层	建立路径或者端到端的连接，通常跨越很长的距离或多个节点	结构化应用：电钮、清水墙
2 数据链接层	将消息集合在一起，添加合适的发送或接收帧头，确保消息在两点之间的传递	基本原则：整个结构的具体支持措施
1 物理层	在物理媒质上传输数据	底层传输实体：为其他阶段（各层）的执行做准备

图 22-3　OSI 参考模型

OSI 模型将任何网络计算设备之间的通信过程划分成了 7 个层或类型；OSI 模型允许数据通信技术开发者和标准开发者之间使用常用的专业术语来商谈两个网络或计算机之间的互联信息，但不能使用所有仅供应商之间的行话。

这些专业术语是 7 层 OSI 模型分层结构的产物，这种分层结构将两个计算机之间的通信任务分割成了各个独立但互相关联的任务，每个任务由其任务层来描述。如图 22-3 所示，顶层（第 7 层）描述了在每个计算机上运行的各个应用程序，因此称为“应用层”。底层（第 1 层）主要描述了两个计算机或网络之间的实际物理连接，因此称为“物理层”。中间的各层（第 2 ~ 6 层）不能很明显地区分成各个独立的层，但是它们充分描述了连接两个计算机的各种不同的逻辑功能组。

为了使用 OSI 模型，网络分析师为每个计算机设备或网络中 7 层 OSI 模型各层对应的网络节点列出了各种已知的协议。这些对应层中的协议集合称为网络节点的“协议堆栈”。例如，使用的物理媒质（如无屏蔽双绞线、同轴电缆或光导纤维电缆）将作为第 1 层的协议加入到协议堆栈中去；而以太网结构或令牌环网络结构将作为第 2 层的协议加入到协议堆栈中去。

OSI 模型允许网络分析师生成一个关于网络节点协议的详细目录，这个目录

记录了每个网络节点的惟一特征，并为网络分析师更深入地研究需要什么样的“协议转换”提供了信息，这些“协议转换”的目的是为了使任意两个网络节点成功通信。最后，为了满足逻辑网络设计的各种需求，OSI 模型为确定物理网络设计中所需的各种硬件和软件技术提供了一个结构化的方法。

与 OSI 参考模型最相似的可能就是大型办公建筑或摩天大楼的设计图了，该图详细说明了分层结构或框架结构的原理。建筑工程中各个承包商最关心的是各层的计划，该计划详细描述了他们的任务规范。但是，每个具体的承包商必须依赖其下层承包商，如同其上层承包商必须依赖他们一样。类似的，OSI 模型的每一层都是独立于其他各层来工作的，但必须依赖相邻层来根据规范协同工作，以完成两个计算机或网络之间的通信任务。

3. 主要类型网络之间的区别

作为自上向下分析中的一部分，计算机或网络节点的地理相邻性是分析信息中的重要组成部分。尽管网络的分类没有很严格的规则，但我们仍然可以列出一些主要的网络类型。

1）远程连接——单个远程用户需要访问本地网络资源。这种类型的组网对于移动专业人员来说非常重要，如销售代表、服务技术人员、现场监察人员等。

2）局域组网——多个用户计算机相互连接以共享应用程序、数据或联网技术，如打印机或 CD-ROM。局域网（Local Area Network，LAN）中可能包含两三个用户，也可能包含数百个用户。LAN 一般只限于建筑物中的单个部门或某一层使用；但是在技术上，所有部门都可以通过一个 LAN 进行联合组网。

3）互联组网——也称为“LAN 到 LAN 的组网或互联”，组网包含了多个 LAN 的连接，在公司中非常常见；在公司中，各部门的用户可以在部门 LAN 上共享数据或进行其他通信。组网面临的挑战是如何在各部门中具有不同协议堆栈（由 OSI 模型确定）的 LAN 之间实现对话，同时只允许经过授权的用户访问互联网络或其他 LAN。各种组网技术还可以将 LAN 连接到主机或迷你计算机上，而不是连接到其他 LAN。

4）广域组网——也称为“企业组网”，涉及到计算机、网络节点或足够远的 LAN 互联，这些远距离的 LAN 需要向电信公司或其他运营商购买广域网（Wide Area Network，WAN）业务。在有些情况中，网络中的广域部分可能是由公司自己所有和操作的。但是，各个节点之间的地理距离是区分广域网的一个决定性因素。广域网的一个子网——城域网（Metropolitan Area Network，MAN）的使用范围为校园或城市区域，其网络覆盖范围通常不会超过数英里。

一个必须记牢的重要事项就是组网的分类是任意进行的，真正重要的是在任何给定的组网机会中选择合适的组网技术（硬件和软件）来实现事先描述的业

务目标。

22.2 个人远程连接

远程连接组网的所有分析与设计方法可以归纳如下：

1）需求分析；

2）逻辑拓扑选择；

3）物理拓扑或分层结构选择；

4）技术审查和详细说明。

1. 远程连接需求分析

远程连接需求分析涉及到远程连接的用户对本地 LAN 资源的性质和使用程度进行记录。这种远程 LAN 连接的逻辑或物理拓扑选择主要取决于远程连接需求分析的结果。远程用户的信息共享需求可能包括以下几个方面：交换 E-mail；上传或下载文件；远程运行互动应用程序；利用 LAN 的附属资源，如打印机；另外，还有一个问题将会直接影响到拓扑的选择，即有多少远程用户需要同时访问本地 LAN 的附属资源？

远程连接分层结构由一个选定的远程连接逻辑拓扑和一个选定的远程连接物理拓扑构成。

2. 逻辑拓扑

远程连接逻辑拓扑通过以下两个方面来区分：应用程序执行的位置（本地或远程 PC）；本地 PC 和远程 PC 之间数据流的性质。最常见的两种远程连接逻辑拓扑或操作模式为：远程客户模式；远程控制模式。

（1）远程客户模式

远程客户模式（也称为“远程访问模式”）通常是指在远程 PC 上执行和保存应用程序，同时只使用共享数据和连接到本地 LAN 服务器的其他本地附属的 LAN 资源。单个本地 LAN 服务器或专用通信服务器可以为多个远程 PC 客户提供服务。网络中，远程节点和任何本地节点具有完全相同的性能。很显然，客户 PC 距离本地服务器是很遥远的。远程 PC 和本地 LAN 服务器之间的数据流量是网络操作系统的数据包，在远程 PC 和本地 LAN 服务器中都安装了该网络操作系统。

（2）远程控制模式

远程控制模式要求为每个远程 PC 指定一个专用的本地 PC，因为应用程序都存储和运行在本地 PC 上的。远程 PC 通过本地 PC 来访问 LAN 服务器的共享数据和其他 LAN 附属资源。远程 PC 只不过是一个简单的输入/输出设备，所有的处理都是在本地 PC 上进行的，在远程 PC 和本地 PC 之间只负责传输输入命令和

屏幕图像；远程 PC 相当于本地 PC 的一个远程键盘和显示器。远程 PC 控制着本地 PC 以及其所有的 LAN 附属资源，因此称为“远程控制模式”。

图 22-4 给出了这两种远程 PC 操作模式或逻辑拓扑的细节、特征和要求。

特　性	远程客户操作模式	Modem 远程控制操作模式
是否需要转向器硬件或软件？	是	否
应用程序实际上在哪里执行？	远程 PC 上	在本地 LAN 附属 PC 上或者通信服务器上
拨号线上数据流量的性质是什么？	LAN 服务器和远程 PC 之间的所有互动信息都在拨号连接上传输	只有来自远程 PC 键盘上的输入数据和输出给远程 PC 显示器的数据才在拨号连接上传输
网络接口	实际上将远程 PC 的序列端口和 Modem 作为一个替代 NIC，这样可以提供一个低速的接口	利用安装在本地 LAN 附属 PC 或通信服务器上的 NIC 以非常高的速度来与 LAN 相接
相对性能	比 Modem 远程控制操作模式慢，产生的数据流量更大，这取决于应用程序是否远程保存或保存在 LAN 上	比远程客户模式快，产生的数据流量更小
别名	远程 LAN 节点	LAN 远程控制

图 22-4　远程 LAN 连接：远程 PC 操作模式

3. 物理拓扑

远程连接物理拓扑是指硬件和媒质的物理排列，媒质可以为远程用户提供访问本地 LAN 的通道。如图 22-5 所示，一个远程 PC 用户可以通过以下 3 种基本的方式来访问本地 LAN 资源：LAN 附属的 PC；通信服务器；LAN Modem。

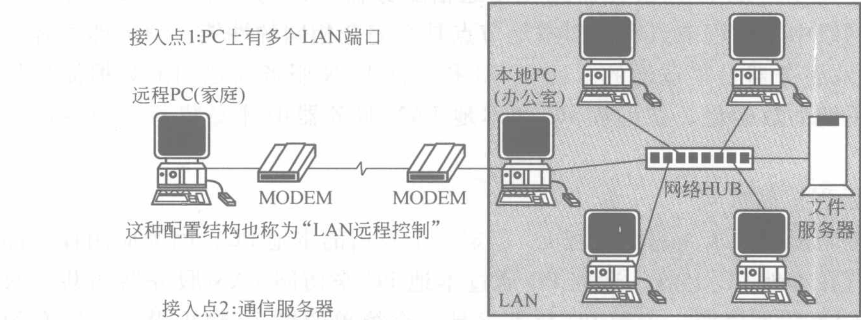


图 22-5　远程 LAN 连接：3 个主要的接入点

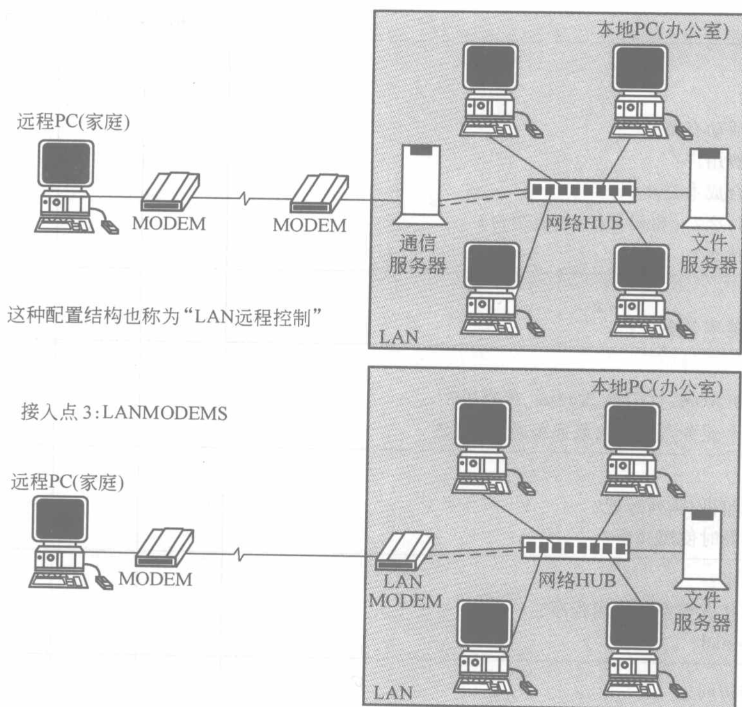


图 22-5 远程 LAN 连接：3 个主要的接入点（续）

有一点很重要，即每个 LAN 访问的实现过程都需要额外的硬件和软件支持，而且这些访问过程可能只限于使用 LAN 的附属资源。

22.3 局域网

图 22-6 列出了局域组网方案中可能存在的业务分析问题，这个列表并没完全给出业务分析问题。任何业务分析问题列表都包含以下两个重要的事项：

1）问题必须深入到所需的信息系统相关的业务活动中。

2）这些问题的答案应该能提供足够深入的认识，以便寻找可能的技术方案。

接下来，我们将主要介绍业务分析问题的分类。

用户满意度是任何成功网络应用的关键。为了使用户满意，首先必须彻底理解他们的需求。除了一个明显的问题之外（即网络必须支持多少用户？），还有很多用于处理个人用户具体业务行为的试探性问题；例如，用户每天会处理很多短小的事件吗？用户需要在每天的某个特定时段传输大型文件吗？是否存在每天某个时段必须完成的活动？这些问题对于确定个人用户的网络通信需求非常重

	当前	2～3年内	5年以内
用户问题： 有多少用户？ 用户的业务活动有哪些？ 什么是预付费用户？ 所有权的综合成本是多少？ 安全需求是什么？（密码保护、超级用户） 支持要点是什么？			
本地通信： 理想的传输速率是多少？			
资源共享： 共享多少CD-ROM、打印机、Modem、传真机？ 服务器与每个业务点之间的最远距离是多少？			
文件共享： 是否需要打印机/队列管理？ 有多少用户同时使用共享？			
应用程序共享： 需要多少和什么类型的应用程序？ 是否需要e-mail？			
分布式数据访问： 共享数据文件保存在哪里？			
LAN管理： 是否需要进行网络管理培训？ 网络使用的难易度如何？			
扩展通信： 网络中有多少MAC？ 需要多少主机连接（什么类型的主机，IBM、DEC、UNIX）？ 是否为内部LAN网络？（LAN到LAN，哪一种网络操作系统（Network Operating System，NOS）？是否需要考虑其他协议？连接是本地的还是远程的？）			

图22-6 LAN和类似LAN：业务分析问题

要。同时，还要实现各种理想的安全等级，因为用户会关心：工资单文件是否也通过网络来访问？谁会访问这些文件，有什么措施来确保授权访问？用户所需的全部技能是什么？是否需要聘用技术人员？能否从外部机构获得支持？

局部通信

我们要记住，局部通信是业务分析问题，而不是技术分析问题，我们不会询问用户网络连接必须有多快。每秒多少位或者每秒多少兆位对于大多数用户来说没有什么意义。如果用户具有如下的业务行为，如计算机辅助设计/计算机辅助

制造（Computer Aided Design/Computer Aided Manufacturing，CAD/CAM）或者其他三维建模或制图软件等那些在线的网络访问行为，那么网络分析师必须清楚存在大量的网络带宽消费者，而且必须记录这些与信息系统相关的业务行为，这些行为可能就是组网带宽的主要消费者。

1. 资源共享

区分哪些资源需要共享以及多少资源需要共享是非常重要的，如打印机、Modem、传真机和这些共享资源的标准位置。共享资源和用户之间的理想距离与可接收的技术选择之间存在一定的关系。

2. 文件共享和应用程序共享

用户需要哪一种程序或软件来完成他们的业务？他们当前正在使用的是哪一种程序？需要购买什么样的新产品？在很多情况下，网络版的软件包其成本比个人PC上的相同软件包的成本低。因此，网络分析师将多个应用程序打包在一起供用户共享。不是所有基于PC的软件包在网络中都是有效的，也不是所有基于PC的软件包都允许多个用户同时访问。一旦将所需的共享应用程序全部打好包之后，必须对这些程序网络版的有效性和性能进行调查，以确保用户满意，并满足业务的需要。

3. 分布式数据访问

尽管我们不能期望用户成为数据库分析专家，但我们必须通过询问大量的问题来确定哪些数据被哪些用户共享，以及这些用户的位置。数据分布分析的主要目的就是确定网络中存储各种数据文件的最佳位置，这个最佳位置通常就是那些接近使用该数据最活跃的用户的位置。在区域型或分支型企业中，通常都共享了一些典型的数据文件，如客户文件、员工文件和资产目录文件。当共享数据的用户超出局域网的范围并且需要通过广域网方式来共享数据时，分布式数据访问就变得非常重要了。网络分析师开始通常会问：各个区域和分支机构是否比较过各种文件格式（这些文件格式用来确定哪些数据需要在整个网络中进行传输）？

4. 扩展通信

局域组网方案在局域网覆盖范围之外的通信能力是各种局域组网方案之间的一个关键区别因素。用户必须清楚地说明LAN之外的需求，如何实现需求是网络分析师负责的事情。扩展通信有时是指与其他LAN之间的通信；如果是这种情况，网络分析师必须调查该目标LAN的所有技术规范，以确定与本地LAN的兼容性。目标LAN可能是本地的（相同建筑物内），也可能是远程的（城镇之外或更远的地方）。LAN到LAN的连接称为“互联组网”，在下一节进行详细介绍。在扩展通信的其他可能方式中，LAN用户必须以本地或远程方式来访问主机。另外，我们只需向用户询问他们需要连接什么，以及这些连接必须在哪里进行；而网络分析师必须说明如何实现这些连接。

5. LAN 管理

在各种 LAN 方案中，另外一个关键区别因素就是网络管理的技巧。如果 LAN 需要一个专职的、经过严格培训过的管理者，那么该管理者的薪水将成为 LAN 成本中的一部分，如同操作成本。其次，用户可能也需要一些管理部件，LAN 方案必须提供这些部件。例如，用户身份的建立或管理，或者对文件或用户目录的访问控制。

6. 预算的真实性

如果组网的成本超出了企业或业务的预算，那么即使最全面、研究最成熟的组网方案也没有任何价值。组网方案的初始研究通常是通过可行性方案报告来描述的，该可行性报告给出了各种可能的网络设计价格浮动范围。高级管理将指出哪些方案值得在经济实用性上进行进一步研究。在有些情况中，高级管理都会有一个大概的项目预算范围，该预算范围可以与网络分析师共享；这个可接受的预算范围（有时称为“每个用户的预算成本”）将作为网络分析师选择技术时的参考。在这种意义上，预算限制就变成了另一个高级业务需求或目标，可以帮助优化最后的组网方案。

7. 规划增长是关键

实际上，用户的需求并不是始终不变的，这些需求会随时间而动态变化。为了设计出不会很快过时的组网方案，必须对用户需求的增长有一个判断力。我们可以想象出网络分析师向管理层解释以下失误时的窘迫，即去年安装的网络无法进行扩展而必须进行替换，是因为无法满足不曾预料到的网络需求增长。一种提高未来组网需求洞察力的方法（见图 22-6）就是向用户询问业务分析中的相同问题，但时间范围变成了 2 ~ 3 年内或者 5 年内。在图 22-6 中，我们难以置信地发现 5 年是给定网络结构或设计的最大规划生命周期。当然，也存在例外，因为终端用户可能不具备制定这些规划时所需的信息或知识。在规划增长和信息需求领域中，网络管理是非常有用的，特别是当一个公司或企业已经进入了正规的策略规划轨道时。

8. 网络应用：LAN 能为你做什么？

除了共享相同的应用软件包（电子表格、文字处理、数据库等）之外（这些软件在联网到 LAN 之前是运行在 PC 上的），组网 PC 还可以提供一些单独的机会来运行一个额外的联网应用程序，这些程序可以明显提高工人的工作效率，或降低成本。

图 22-7 归纳出了 LAN 常见应用的属性和要点。我们必须注意，图中的用户、特征和要点（接下来将要介绍）只适用于图中列出来的各个局域网应用功能。很多相同的功能在多个 LAN（互联组网）或远程 WAN 上运行时，就会变得非常复杂。下面我们将简要介绍每一种 LAN 应用。

网络应用	要点
网络打印	共享打印机、打印机需求、缓冲和联机外部操作；第2代网络打印机更快/更小；排队/设备管理、多种打印格式
网络备份	配置硬件和软件以满足需求；检查存储特征；规划独立的备份；日志报告的生成；硬件相关性和容量
网络安全	病毒控制、额外的用户ID/密码保护、加密、用户授权、不活跃终端的自动注销
网络管理	网络显示、诊断和分析工具，这些工具只与电缆和网络类型有关；管理事项；NOS中的简单LAN软件
网络资源优化	硬件和软件目录数据库、报告和询问能力；规划和网络配置工具；监控和网络报警能力
配置管理与目录管理	
远程访问控制软件	包含在很多完全集成的LAN中
连接性/网关、软件、MAC集成软件	
组件	工作流程自动化、协同工作、文档审定、信息共享、团体调度、增强型电子邮件

图22-7 LAN和类似LAN：网络应用分析

(1) 网络打印

随着用户需求和技术的不断发展，网络打印技术也在不断发展。

在一个典型的LAN上，一个联网的PC可能会通过网络接口卡在网络上发送一个打印请求；这种联网打印的业务请求会被一个设备接受，该设备负责组织网络打印机来响应打印请求。根据LAN的不同配置，该设备可能是一台连着打印机的PC、一台连着打印机的专用打印服务器或者直接是一个联网的打印机。该打印过程中，必须有某个软件来专门管理请求、假脱机、缓冲、排队和打印过程；这个软件可以是全部网络操作系统的一部分，也可以是专门为网络打印机管理而编写的程序。

(2) 网络备份

在网络上备份数据和应用文件，对于整个网络安全和从不可避免的数据破坏灾难中恢复来说是必需的。尽管备份过程及其全部构成相对非常简单，但实现起来就不是那么容易了。网络备份系统基本上由两个部分构成：软件（用来管理备份）；硬件（用来记录备份文件）。

有一些网络备份软件和硬件只能在特定的网络操作系统中一起工作，其他的网络备份软件只能与其配套出售的硬件一起工作。硬件设备和软件（如操作系统或网络操作系统）之间的相互协作通常是由专用软件程序（驱动）来控制的。

我们必须很清楚，磁带备份设备或软件提供商必须提供必要的驱动，以确保磁带设备的可操作性。硬件设备可以是各种类型的磁带子系统或光学驱动器，这些硬件设备之间的主要区别包括：

1）价格是多少？备份设备的存储容量是多少？

2）运行速率有多快？数据能以多快的速度从 PC 传输到备份设备？如果你正在备份大容量磁盘驱动器，这项属性非常重要。

3）压缩率是多少？数据能否以压缩格式存储在备份设备上？数据如果压缩过，将可以在备份介质上节省大量的空间。

记住，备份不是一个必须进行的过程。备份文件的保存及其实现难易度才是备份过程考虑的要点。对于网络管理者来说，规划自动备份的能力和打印备份/保存活动日志报告的能力同样重要。

（3）网络管理

网络管理的全部任务通常被划分成至少 3 个不同的操作过程。首先，必须对网络进行监控。这样，网络管理软件的主要任务之一就是监控 LAN，以检测任何非常规的活动，如网络适配卡产生故障，或者一个罕见高数据传输速率，该速率垄断了各工作站之间共享的全部带宽。先进的 LAN 监控程序可以将网络的地图显示在图形终端上，操作人员可以对某个节点或工作站进行放大，来获取更加详细的信息和性能指标。有些监控程序还可以将当前网络活动与预先设定的可接受参数进行比较，当网络活动出现异常时，可以通过点亮屏幕上的节点标志红灯来进行报警。监控软件也是为 LAN 上的监控文件服务器专门编写一段程序。除了监控服务器性能和启动警报之外，有些监控软件还可以拨号，并寻呼那些远离网络控制台的网络管理人员。

一旦某个问题被监控到并识别之后，必须对其进行分析和诊断，这就是网络管理软件的第二个主要任务。诊断通常是通过一系列的设备进行的，如协议分析仪或探测仪；这些设备附着在 LAN 上，负责监视、测量、记录每一个通过的数据位。通过在 LAN 上不同的点部署多个探测仪（或者称为“分布式探测仪”），我们就可以发现瓶颈并找到导致性能减退的原因。LAN 测试设备必须能测试并隔离 LAN 中的以下 3 个主要部分：网络的线路或光缆；网络适配卡（光缆与工作站之间的接口）；产生网络活动的工作站或 PC。

诊断出产生组网问题的原因之后，必须采取正确的措施来解决这些问题。例如，性能失常的网络接口卡必须废除，而工作站上垄断网络资源的应用程序也必须终止。采取措施的行为就称为网络的“管理”。这样，“管理”就与常用的“网络管理”之间形成了对比。通常，用来管理 LAN 的理想管理软件包含在网络操作系统之中。

对于大多数网络操作系统来说，LAN 监控软件与其他专用网络管理功能只

是一种副产品。如果这些副产品不是由原始的网络操作供应商生产的，而是由另外一个公司生产的，那么这些副产品就称为“第三方产品”。这些第三方产品通常具有很高的质量，但在购买时应更加谨慎，相关软件或网络操作系统未来版本的兼容性没有保证。

（4）网络安全

除了典型的安全特征（如大多数操作系统都提供的密码保护、目录访问控制）之外，LAN中使用了更加先进的网络安全软件/硬件。例如，安全软件可以添加到工作站或服务器，并实现以下功能：

1）在PC导入之前，需要用户身份和有效密码才能进入；

2）对重要数据或应用文件进行加密，以防止被篡改；

3）自动注销不活动的终端，以阻止未授权的终端访问系统资源；

4）允许特定的用户运行特定的应用程序；

5）通过安全鉴定程序对用户的真实性进行鉴定。

人们关注的另一个网络安全领域就是病毒。病毒控制软件有时包含在网络安全数据包中，病毒控制实际上分为3个阶段，高效的病毒控制软件应该至少着重于以下3个方面：

1）病毒防护：充分控制用户对系统的访问，以防止未经授权的用户感染LAN；

2）病毒检测：无论病毒伪装得多好，先进的软件都可以发现病毒；

3）病毒清除：有时称为“抗病毒程序”，这种软件可以清除病毒的所有踪迹。

(5) 组件

组件是一类软件的统称，该软件吸取了一个经验，即所有的工人被联网在一起来最大程度地提高工作效率。组件通常描述的是以下的软件类型：工作流程自动化、协同工作、团体调度、文档审定、信息共享、电子白板和增强型电子邮件。

9. 局域网体系结构

网络结构的选择将会对网络适配卡和媒质以及网络操作系统的选择产生或多或少的影响。例如，以太网结构需要以太网适配卡。在稍后的内容中我们就可以发现，正是适配卡或MAC层协议决定了网络是以太网结构还是令牌环、光纤分布式数据接口（Fiber Distributed Data Interface，FDDI）结构，或者其他网络结构。以太网运行在各种粗细同轴电缆、屏蔽或无屏蔽的双绞线、光纤或无线媒质上；因此我们可以看出媒质的选择范围很广。

10. 以太网

以太网（遵守IEEE 802.3标准）是一种基于冲突检测的载波侦听多路接入

(Carrier sense multiple Access with Collision Detection, CSMA/CD) 网络结构；过去通常位于总线配置中，但现在大多数位于基于HUB的星形拓扑结构中。每个附着在以太网上的设备（多数为网络适配卡）具有惟一的硬件地址，该地址是以太网在构建时分配的。当新的设备添加到以太网中时，其地址对于其他以太网设备来说就成了新的目的地址了。

以太网的媒质接入层协议原理根据标准格式来形成数据包，并在共享媒质上进行传输，这种以太网数据包格式与IEEE 802.3标准中的数据格式几乎完全相同，因此这两种数据包格式经常被交替使用。

以太网中的冲突和重发隐患是由以太网的CSMA/CD接入方法带来的。在有些情况中，一个拥有100~200个用户的以太网几乎感受不到网络的容量。但是，传输数据的性质是决定网络潜在性能问题的关键因素。相比图形用户接口(GUI)面向屏幕的数据传输（如基于Windows的应用），基于特性的传输产生网络容量问题的可能性就小得多（其中，少许特性被输入并在网络上发送，如典型的数据登录）。CAD/CAM图像对带宽的要求更高。如果30个或更多的工作站同时请求基于Windows的全屏数据传输时，在以太网上将会造成冲突和网络容量问题。对于任何数据通信问题，都有相应的解决方案。解决这些问题的关键点就是提供一个保证措施，即尽管以太网的网络容量是无限的，但在大多数情况下，必须提供足够的带宽。

11. 令牌环

IBM的令牌环网络结构（遵守IEEE 802.3标准）使用了一个星形拓扑结构、连续消息发送机制和令牌通行访问方案。由于连续逻辑拓扑相当于令牌消息从相邻的节点开始传输，经过一个环后，再回到相邻节点，因此令牌环网络结构有时也称为"逻辑环"或"物理星"。令牌通行访问法中的令牌环为这种网络结构提供了一项关键性的优势特性。通过令牌通行访问法可以确保数据传输时不产生数据冲突，这一点非常适用于那些需要保证传输可靠性的应用环境。

12. FDDI

随着网络结构的不断发展，越来越多的用户加入了LAN，因此网络的整体带宽需求也不断增加。LAN不仅在规模上增加不少，而且在复杂性上也增加了很多。各种协议的LAN通过网桥和路由器进行联合组网意味着将会产生更大的数据流量。各种网络应用正在驱动着带宽增长需求的不断发展。分布式计算、数据分布和客户/服务器计算等等，所有这些概念都依赖于高带宽和高可靠性的网络结构基础。成像、多媒体和数据/语音合成等技术都需要很高的带宽来实时传输和显示各种不同的数据格式。换句话说，如果LAN中需要传输全移动视频（作为多媒体程序的一部分），必须有足够的带宽来支持视频以全速运行，而不会出现停顿。类似的，数字化语音在具有足够带宽的LAN中传输时，必须听起

来很正常。

当遵守基于标准的协议时，FDDI不仅可以提供很高的带宽，而且还可以提供很高的可靠性和安全性。FDDI的可靠性不仅仅来自于光纤自身的优势，因为我们知道光纤可抗电磁干扰（Electro Magnetic Interference，EMI）和射频干扰（Radio Frequency Interference，RFI），而且通过FDDI的物理拓扑设计也可以提高FDDI的可靠性。

FDDI的物理拓扑是由两个环构成的，数据在传输时同时沿着两个环的不同方向进行传输。其中某一个环是主数据环，另一个是次数据环或备份数据环，只有当主环出现故障时才会使用次环。但是这两个环都附着在一个单一的HUB或集线器上，HUB中的单点故障将会造成网络媒质出现数据冗余现象。除了速度和可靠性之外，距离也是FDDI LAN的一个重要特征。

FDDI的另一个优势特征就是它可以很轻松与以太网兼容。这样，一个业务流就不必为了升级到FDDI而废弃它现有的整个网络，只需要在FDDI网络和以太网之间添加一个网桥就可以了。

FDDI网络结构的应用主要分为以下3种类型：

1）校园骨干网：不必是一所真正的大学校园，这种结构用来连接那些距离很近的LAN；

2）高带宽工作组：当FDDI LAN真正作为局域网使用时，第二种应用类型就是连接PC或者连接那些需要高带宽来相互通信的工作站。多媒体工作站、工程工作站或CAD/CAM工作站就是高带宽工作站的典型例子；

3）高带宽子工作组连接：在某些情况下，只有2~3个设备（可能是3个服务器）具有高带宽需求。随着分布式计算和数据分布的增加，服务器到服务器之间的高速数据转移需求也在不断增加。

13. 无线LAN

前面介绍过的很多网络结构都是基于一种或多种物理媒质的，而另一种传输媒质选择方案就是局域网的无线传输，无线传输不需要任何物理媒质。在局域网技术领域，目前主要有3种主流的无线传输技术，分别是：微波传输、扩频传输和红外传输；它们都是不同频率下的无线电传输技术。

14. 无线LAN的应用

无线LAN的各种主要应用降低了无线技术的接入难度。便携式或笔记本式PC之间可以利用自身的无线LAN适配器来构建一个临时的LAN连接，只需要PC处于基于服务器的无线LAN适配器或无线HUB的覆盖范围之内即可。这样，学生或员工就可以通过在便携式PC中安装一个合适的无线适配器，在无线HUB覆盖范围内的任何地方坐下来，加入到一个LAN中。

会议室也可以通过装备无线HUB来为各个自由的工作组提供无线接入服务，

而无需在会议室内布置电缆。类似地，通过快速安装无线 HUB 和带有无线适配器的便携式 PC，就可以满足临时扩容的需要，或者很方便地用于紧急事务/抢险救灾中。在无线 LAN 中，无需重复布线，也不用查找线路中的交叉点，这一项优势非常重要。

总之，无线 LAN 技术允许整个 LAN 进行重新配置，以方便从一个地方转移到另一个遥远的地方；远处的非技术性用户只需要将电源线插入相应的接口，他们就拥有了一个临时的 LAN。对于拥有大量远程子公司而技术人员不多的公司来说，无线 LAN 技术是最理想的选择。无线 LAN 无需预先安装站点，而且还可以避免为构建有线 LAN 和进行有线故障维护而产生的消耗和管理成本。

局域网硬件

1. 服务器

服务器的工作就是管理联网资源在各个客户 PC 之间的共享。根据客户 PC 的数量以及共享资源的等级，有必要设置多个服务器或指定某些服务器来专门管理资源类型。如果服务器变成了专用服务器，那么共享资源必须以某种逻辑方式进行分组，以便优化服务器在管理某种类型资源共享时的性能。潜在可共享的网络资源包括：文件、应用程序、数据库、打印机、对其他 LAN 的访问（本地的）、对其他 LAN 的访问（远程的）、对信息业务的访问以及远程 PC 对 LAN 的访问。

2. HUB/多站接入单元

星形物理拓扑（可以使用以太网和令牌环）的中心就是配线中心，也称为“HUB”、“集线器”、“转发器”或“多站接入单元（Multistation Access Unit，MAU）”。

（1）转发器

转发器，顾名思义，就是复制它接收到的每一个数据。这种复制行为实际上是在数据从一个附着设备或 LAN 分段传输到另一个附着设备或 LAN 分段之前，对数据进行重新记时和重新生成。

（2）HUB

“HUB”、“集线器”或“智能集线器”这些词语通常交替使用。尽管无法阻止制造商根据他们的喜好来选择称呼，但是我们仍然可以区分这两大类的配线中心。“HUB”通常用来描述具有固定数量端口的单机设备，该设备不仅仅是一个转发器。媒质连接的类型和 HUB 提供的网络结构在制造时就确定了。例如，一个 10BaseT 以太网 HUB 可以为以太网提供固定数量的 RJ-45 双绞线连接。HUB 通常不支持其他类型的媒质类型或网络结构。

3. MAU

MAU 是 IBM 公司对令牌环 HUB 的称谓。MAU 在制造时可以包含固定数量

的端口、无屏蔽双绞线连接和屏蔽双绞线（Shielded Twisted Pair，STP）连接。IBM公司在屏蔽双绞线与MAU的连接上采用了专用的令牌环连接器。MAU可以提供不同程度的管理能力。

（1）集线器

“集线器”或“智能集线器”通常用来描述灵活性和可扩展性突出的设备。集线器的基座通常是一个盒状的空心设备，称为“底盘”；该底盘中包含了一个或多个电源以及一个内置的网络中枢。该网络中枢可能是以太网、令牌环、FDDI或几者的综合体。在该底板中，还插入了各种单独的插件和各种模块。

例如，一个8-端口或16-端口的双绞线以太网模块可以插入到集线器的底盘中；支持SNMP（Simple Network Management Protocol，简单网络管理协议）的网络管理模块可以插入到靠近先前安装的10BasesT端口模块的底盘中。在这种混合搭配的配置中，可以添加附加的插件来连接PC和令牌环适配器或PC，或者连接工作站和FDDI适配器或其他无声异步终端。

通过媒质接口与PC或工作站上网络接口卡之间的双绞线连接，工作站可以连接到一个网络集线盒上。值得注意的是，以太网可以运行在非屏蔽双绞线（Unshielded Twisted Paire，UTP）、STP、粗细同轴电缆以及光纤上。

附加的模块并不是对所有的集线器都适用，集线器可能允许某个网络中的数据流通过网桥或路由器上的附加模块来传输到其他本地LAN（网桥和路由器将在下一节中介绍）。这种组合的集线器有时也称为“互联组网HUB”，通过添加其他专用插件或模块，就可以实现与远程LAN或工作站的通信；这些专用的插件或模块可以提供与广域网设备的接入端口，广域网设备可以从一般的网络运营商处购买。

（2）交换HUB

集线器和HUB提供的网络集线盒缩短了网络中枢的长度，但并没有改变网络中枢的结构特征。例如，在一个以太网集线器中，多个工作站通过各种媒质来访问一个内置的以太网中枢，但是以太网的基本算法（如CSMA/CD访问法）仍然控制着该以太网在集线盒中的性能。在某一时刻，只能有一个工作站在共享的中枢上广播消息。

交换HUB就是用来解决这个问题的，即某一时刻只能由一个工作站广播消息。交换HUB可能会产生数据冲突、数据重发，而且可能减小高带宽需求设备之间的吞吐量，如工程工作站或服务器到服务器的通信。

以太网交换机实际上可以在数据包交换的基础上，在任意连接到以太网上的两个设备之间创建连接或交换；这样，通过一个以太网交换机就可以解决前面提到的某一个时刻的广播限制问题了。

集线器技术分析

图 22-8 给出了用来比较分析的主要技术特征。在购买一种 HUB 之前，我们可以参考以下这些性能特征。下面归纳了在购买 HUB 或集线器之间必须仔细考虑的几个主要标准：可扩展性；支持的网络结构；支持的媒质类型；可扩展通信能力，即终端支持、互联组网的选择、广域组网的选择；HUB/集线器管理能力；可靠性特征。

特性/特征	说明/含义
可扩展性	内容：端口、LAN、扩展槽 注意事项：大多数单机 HUB 是不能扩展的，即使可能串联起来 集线器可以随着它的整体扩展（多个开放式扩展槽）或而产生变化；或者通过添加额外的端口来实现扩展；多个集线器只需要一个 LAN 中枢模块，你是否需要在 2～3 年或 5 年之内对集线器进行扩展
网络结构	内容：以太网、令牌环、Arcnet、FDDI、AppleTalk 有些集线器只适用一种类型的 LAN 模块，以太网广泛适用各种类型的 LAN 模块；在购买集线器之前，确定是否支持所有必需的 LAN 类型
媒质	内容：无屏蔽双绞线、屏蔽双绞线、细同轴电缆、粗同轴电缆、光纤 注意事项：每种不同的媒质必须连接到具有不同类型接口的 HUB 上；确保端口模块可以为媒质的附加模块提供正确的接口，这种集线器/HUB 与你的网络适配卡是否匹配
扩展通信：	
支持 Macintosh 计算机	Apple 公司的 Macintosh 型号计算机能否连接到 HUB 上
支持终端	支持无声异步终端直接连接的模块是否有效？接口是什么（例如，DB-25、RJ-45）
互联组网	内容：以太网、令牌环、Arcnet、FDDI、AppleTalk 互联组网方案是惟一的，该方案取决于两个 LAN 的连接方式；换句话说，除非集线器提供一个专用模块（该模块专门负责连接以太网模块和令牌环 LAN），集线器中的互连通信才能进行
广域组网	内容：与不同承载设备的接口——DDS、DS-0、T-1、X.25、帧中继、SMDS、ATM
管理	内容：协议——SNMP、CMIP、CMOT 可以提供哪些管理等级 1. 各个端口能否进行管理 2. 是否包含监控软件 3. 广域和局域连接的性能能否通过集线器进行分析/监控 4. 一般可以提供什么样的安全等级和专门的管理功能 5. 端口能否远程激活和终止 6. HUB 和集线器能否通过附加的工作站或 Modem 来控制 7. 图形用户接口能否提供管理和监控功能 8. 报警门限值如何设置 9. 故障如何处理 10. 端口访问能否受日期和时间限制 11. 管理软件运行在什么样的操作系统上，DOS、OS/2、Windows NT、UNIX、APPLE 12. 能否显示网络结构图
可靠性	内容： 1. 是否包含集成的 UPS 2. 是否存在多个、多余的电源 3. 单个模块能否进行热插拔

图 22-8 集线器技术分析

1. 网络接口卡

网络接口卡（也称为“网络适配卡”）是客户或服务器PC与网络共享媒质之间的物理连接设备。在网络和PC或工作站之间提供这种接口时要求网络适配卡必须能够遵守网络结构的接入规则。这些由网络适配卡实现的软件规则，通常称为“媒质接入控制（MAC）”协议，该网络适配卡控制着与共享网络媒质的接入；MAC协议包含在OSI七层参考模型中数据链路层（第2层）的MAC子层中。

由于MAC层接口卡和MAC层接口协议的存在，我们可以认为正是适配卡决定了网络的结构以及各种要素协议，而不是其他的组成部分。如果将PC扩展槽中的以太网适配卡拔出来，并替换上一个令牌环适配卡，那么就可以得到一个令牌环工作站了；同时，物理媒质也不需更换，因为以太网、令牌环和FDDI/CDDI技术通常是工作在相同的物理媒质上。

2. 适配卡驱动器的作用

购买的适配卡通常可以成功地接入到CPU的总线上，而且网络结构的媒质可以确保硬件的连接性。但是，完全的互操作性取决于选择的网络适配卡，该适配卡可以与网络操作系统和安装该适配卡的PC操作系统之间成功地进行通信。

22.4 互联组网

应用：互联组网具有什么功能？

目前互联组网有一个必然的发展趋势，即任何组织机构很快将通过多种信息平台或信息体系来共享各种信息。这种信息共享可能存在于各个LAN之间，与网络结构或网络操作系统之间的信息共享不同。合并了多个计算平台或各种网络结构的信息系统和网络操作系统通常称为“企业数据处理技术”环境；在“企业数据处理技术”环境中运行的就是“企业网络”或“企业互联网”。进行成功联合组网的关键是，对于LAN上的工作站终端用户来说，与企业数据处理资源之间的连接必须完全透明。

换句话说，一个终端用户不必知道数据库服务器或者被访问的磁盘驱动器的具体物理位置。所有用户只需知道节点名称或驱动指示器的字母，以及这些节点名称或驱动指示器字母什么时候输入；在访问数据时，用户不用关心这些节点或驱动器的具体位置，因为它们可能位于某个房间，也可能位于另一个国家。

随着人们对信息重要性认识的提高，很多公司已经将信息技术看做是提高公司竞争优势的一种无形资产。在正确的时间、以正确的格式将信息传输到正确的地点就是信息系统和互联组网的目标。商业团体、接管企业和合伙企业已经加速提出了对穿越地理或物理LAN边界来无缝共享信息的需求。智能互联LAN设备

负责掌握网络附属资源的位置并将数据包从一个 LAN 传输到另一个 LAN，这个任务更加复杂，因为不同的 LAN 在以下几个方面各不相同：网络结构；媒质；网络操作系统；操作系统。

LAN 之间的这些区别是由各种规则或协议定义的，这些协议还定义了网络通信过程的各个方面。当需要共享信息的 LAN 根据不同协议来协同工作时，必须使用具有协议转换功能的互联 LAN 设备；这样，互联 LAN 设备就可以穿越逻辑边界（与物理或地理边界相反）来传输数据，这些互联 LAN 设备包括转发器、网桥、路由器和网关。尽管严格来说这些设备的功能在技术上有所区别，但是不能保证制造商就会遵从这些区别。数据设备的功能性分析是确保互联 LAN 设备满足互联组网连接需求的最好方法。目前，互联组网分析和设计是一个非常复杂和混乱的研究领域。本节的主要目标就是帮助读者熟悉互联组网技术和应用，并为下一步的研究提供资料。

OSI 模型和互联组网设备

OSI 模型首先作为一种常见的框架结构或参考模型在 22.1 节中作过介绍，该模型中的协议和组网系统可以提供一定的互操作性。在本节中，OSI 模型可以提供一个高效率的框架结构，该框架结构中的操作特性与前面介绍的互联组网设备的操作特性有所区别。图 22-9 描述了每种互联组网设备与其相关 OSI 层之间的关系。接下来，我们将详细介绍每种设备和相关的 OSI 模型分层。

图 22-9　OSI 模型和互联组网设备之间的关系

互联组网技术

1. 转发器：第 1 层——物理层

我们知道，LAN 上的所有数据流量是指在某个离散时间点上运行在物理媒质上的数字离散电压。因此，转发器的任务就很容易理解了，如下所述：

1）通过重新生成和计时输入信号来复制数字信号；

2）在所有的附属分段传输所有的信号；

3）不会读取数据包的目的地址；

4）允许各种不同媒质的连接；

5）通过复制 LAN 分段间的信号来有效扩展 LAN 之间的传输距离。

一个转发器就是一个对数据一视同仁的互联组网设备，每个信号从转发器的

一端输入，经过复制后，从另一端输出。转发器既适用于以太网结构，也适用于令牌环网络结构，同时适用于各种类型的媒质。转发器是一个物理层设备，因此它只关心与信号电压和时序相关的物理层信号协议。上层协议很难严格区分，如以太网和令牌环帧协议（第2层，数据链路协议）。因此，转发器在生产时必须指明是用于以太网结构还是用于令牌环结构。转发器的功能主要包括：通过复制多个LAN分段的信号来提高网络媒质的传输范围；隔离不同LAN分段的重要网络资源；部分转发器还允许在不同类型的媒质上互联相同网络结构的网络分段。

2. 网桥：第2层——数据链路层

(1) 网桥的功能

网桥的主要应用场合与时机包括：LAN分段上的网络流量已经达到性能允许的极限；对部门LAN到公司LAN中枢的访问进行控制，以便本地LAN数据不会在公司的中枢网络中产生不必要的拥塞问题。

通过对网桥连接上多个LAN分段的用户进行分组，可以降低每个LAN分段上的数据流量，同时还可以按照逻辑方式对用户进行分组。在对用户进行分组时，必须根据任务功能、相互之间的通信需求和对服务器上存储数据的访问需求来进行划分。这样分组之后，LAN中80%的数据流量保留在本段LAN中，只有20%的数据流量需要穿过网桥达到相邻的LAN分段。

通过网桥来控制对公司中枢网络的访问，可以保障企业网络的通信性能，即在公司的网络中枢上只允许基本的网络通信。服务器和其他联网设备可以直接连接到公司的网络中枢上，同时所有用户的工作站可以连接到各个LAN分段上，这些LAN分段通过网桥进行隔离。

当一个LAN上的用户偶尔需要访问另一个LAN上的数据或资源时，就需要一个互联组网设备，该设备通常比转发器更加先进，更加有判别能力。通过比较网桥和转发器之间的功能，我们可以发现网桥的判别能力更强。网桥不仅可以转发所有LAN或LAN分段之间的数据，还可以读取每个数据帧的目的地址，并确定该地址是本地的还是远端的（网桥的另一边），同时只允许那些不是本地的目的地址数据帧通过网桥到达远端的LAN。

那么网桥是如何知道目的地址是不是本地的呢？这是由于数据链路协议（如以太网）在预先定义的以太网帧方案中就包含了源地址和目的地址。网桥还可以检查它接收到的每个数据帧的源地址，并将该源地址添加到已知的本地节点列表中。当读取完一个目的地址之后，网桥就会将该目的地址与本地节点列表中的地址内容进行比较，来确定该数据帧是否需要通过网桥（即判断目的地址是否为本地地址）。因此，网桥有时也称为“非本地则通过”设备。

上面的读取、处理和判别过程显示了网桥的先进性能，这些性能是由安装的软件提供的。

(2) 网桥的分类

网桥包含很多类型。实际上，网桥可能是插在PC扩展槽中的一个插件，也可能是一个独立的设备。尽管我们知道网桥在两个LAN之间可以进行互联操作，但是这个操作的性质以及网桥的输入和输出接口是由该网桥连接的两个LAN的特性决定的。在决定输入和输出网桥的属性时，我们必须注意以下要点：MAC子层协议、LAN的速度、本地或远端、广域网设备和媒质；这些要点分别描述如下。

1）MAC子层协议：该协议取决于MAC子层或桥接LAN的网络结构，同时需要以下任何一种网桥：透明网桥、转换网桥、密封网桥、源地址路由网桥、源地址路由透明网桥和自适应源地址路由透明网桥。首先，确定进行桥接的两个LAN是以太网结构还是令牌环结构。连接相同数据链路格式LAN的网桥称为"透明网桥"。特殊类型的网桥包括"格式转换器"，用来桥接以太网和令牌环网；这些特殊网桥有时也称为"多协议网桥"或"转换网桥"。第三类网桥（类似于转换网桥）用来桥接以太网和FDDI网。与转换网桥不同的是（转换网桥在对数据进行重新封装之前，必须对数据链路层消息进行处理），密封网桥只是将整个以太网数据链路层消息封装成一个数据包（数据帧），该数据包遵守FDDI数据链路层协议。"源地址路由网桥"专门用来连接令牌环LAN。可以支持源地址令牌环LAN与非源地址路由LAN连接的网桥称为"源地址路由透明网桥"。最后，可以支持各个透明桥接以太网LAN分段之间相互连接、源地址路由令牌环LAN分段之间相互连接，以及这两种情况组合连接的网桥称为"自适应源地址路由透明网桥"。图22-10给出了这些不同类型网桥的具体描述。

网桥类型	LAN链路分段
透明网桥	·以太网到以太网 ·非源地址路由令牌环到非源地址路由令牌环
转换网桥	·以太网到令牌环、令牌环到以太网
密封网桥	·以太网到FDDI、FDDI到以太网
源地址路由网桥	·源地址路由令牌环到源地址路由令牌环
源地址路由透明网桥	·源地址路由令牌环到以太网、以太网到源地址路由令牌环
自适应源地址路由透明网桥	·以太网到以太网 ·源地址路由令牌环到源地址路由令牌环 ·源地址路由令牌环到以太网、以太网到源地址路由令牌环

图22-10 不同网桥与LAN分段的链接

2）LAN的速度：我们必须清楚输入LAN和输出LAN的速度，以确定网桥需要进行什么样的速度转换。

3）本地或远端：在确定了MAC层协议和LAN的速度之后，必须仔细考虑各个LAN之间的地理相邻性。如果两个LAN之间的距离不能满足传统LAN媒质（如UTP、同轴电缆或光纤）链接时的要求，那么必须使用网桥，该网桥必须提供合适的广域承载业务接口。

（3）网桥的性能

网桥的性能通常以下面两点作为衡量标准：

1）过滤率：衡量标准为每秒多少数据包或者每秒多少帧。网桥在以太网帧或令牌环数据包上面读取到目的地址后，将确定哪个数据包应该通过网桥接入到互联网，这个过程称为“过滤”。

2）发送率：衡量标准也是每秒多少数据包或者每秒多少帧。在过滤过程确定了哪个数据包应该通过网桥接入到互联网后，网桥必须执行一个隔离操作，即将数据包发送到互联网媒质上去，无论该数据包是到达本地的还是远端的。

（4）网桥、协议及OSI模型

网桥负责读取数据帧中的目的地址，该数据帧的格式遵循预先定义的结构或协议。换句话说，以太网和令牌环网络结构为数据帧的格式定义了一个“bit构成（Bit-by-Bit）”协议，因此，网桥根据这个协议就知道了应该在哪里检查以太网数据帧来查找目的地址位。按照OSI模型的定义，以太网和令牌环都是MAC子层协议；MAC子层是OSI模型中第2层（数据链路层）的两个子层之一。另一个数据链路子层称为“逻辑链路控制”子层。由于网桥读取和处理数据的协议位于MAC子层中，因此，网桥有时也称为“MAC层网桥”。

嵌入在以太网帧数据字段中的全部都是OSI高层协议，这些高层协议的变化与以太网数据链路层协议无关。换句话说，数据链路层协议（如以太网和令牌环）就是“网络结构”，而网络层协议来自不同的网络操作系统。网桥只需要关注网络结构（MAC子层）的协议或格式，对上层协议可以完全忽视不管。

大多数网络操作系统实际上是由各个协议栈构成的，而每个协议栈是由OSI模型的3~7层中每层的独立协议组成的。网络操作系统中的每一个协议都具有不同的相关组网功能，这些功能与OSI模型中对各层的功能性定义相对应。例如，网络层协议TCP/IP套件就称为“互联网协议（IP）”。

3. 路由器：网络层处理器

数据包在传输到目的地址之前会经过多个LAN或广域网链接，完成这个任务的设备就是“路由器”，路由器主要用于以下情形：

1）构建大型分级网络：路由器用来生成中枢网络。

2）参与或为大型分级网络（如Internet）提供接入。

（1）路由器功能

尽管路由器和网桥都可以检查和发送数据包，但它们主要在两个方面有所区

别。首先，虽然网桥可以读取LAN上每个数据包的目的地址，但路由器只检查那些地址指向该路由器的数据包；其次，路由器不仅可以允许数据包以类似网桥的方式接入到互联网中，而且更加谨慎和高效。在对数据包的发送进行区分之前，路由器首先确认目的地址的存在和到达目的地址的最新有效网络路径信息。接着，根据最新的流量情况，路由器将选择最佳的数据包传输路径，并发送数据。这个最佳路径是相对的，它主要由各种不同的协议控制，而且最佳路径的检查非常快。

路由器本身也是一个目的地址，可以在直接连接到网络或间接连接到网络的任何地方接收、检查和发送数据包。以太网数据包或令牌环数据包上的目的地址必须是路由器的地址，该路由器将进一步处理互联网上的数据。这样，数据链路层目的地址字段就会对该路由器进行描述；然后，该路由器就会丢弃该MAC子层封装（包含了子层地址），进而读取以太网帧或令牌环帧中数据字段中的内容。

如同数据链路层协议一样，网络层协议也定义了一个“bit构成（Bit-by-Bit)”数据帧结构。路由器将对外部特征像数据的数据包和被数据链路层组网设备（网桥）忽视不管的数据包进行分解并彻底检查以确定下一步的处理。路由器在读取了网络层目的地址和网络层数据协议之后，就会查询“路由表”来确定发送数据包的最佳路径；在找到最佳路径之后，路由器就会根据最佳传输路由对数据包进行重新封装。

举例来说，如果选择一个分组交换数据网络用于广域链路的传输，那么本地路由器将会把数据包压缩到合适的封装中。另一方面，如果最佳路径位于本地以太网连接中，那么本地路由器将会把数据包放回到一个新的以太网封装中，并在该数据包原来的路径上进行发送。

和网桥不同（网桥只允许接入到互联网(“非本地则通过”逻辑)），路由器特意使数据地址指向了远端路由器。但是，在路由器实际将数据包释放到互联网上之前，路由器必须确认数据包的目的地址存在。只有当路由器同时满足目的地址可行性和意向路径的质量时，它才会释放那些经过仔细封装的数据包。路由器上这种谨慎的处理行为称为“确认是远端目的再发送”逻辑。

（2）最佳路径的确定

最佳路径必须考虑以下因素：

1）中间节点的跳数，即在到达目的地之前，数据需要经过多少个其他路由器的处理？每个路由器处理数据的过程都会耗费一定的时间，因此，路由器越少，数据的整体传输就越快。

2）通信电路的速度或环境。路由器可以动态维持路由表，使其与网络流量信息保持一致。

3）以太网封装中可以包含多个协议，如网络操作系统协议。我们可能会要求路由器打开以太网封装并将所有的NetWare（IPX）流量发送到某个网络，同时将所有的TCP/IP（IP）发送到另一个网络；这种情况下，某个协议就需要进行优先级处理。

（3）多协议路由器

路由器通过读取指定的网络层协议来提高过滤率和发送率。如果一个路由器只负责发送一种类型的网络协议，那么该路由器必须非常清楚应该在哪里寻找目的地址，并可以更快地处理数据包。但是，考虑到不同的网络协议具有不同的数据结构，而数据结构中目的地址的长度和位置又各有区别，因此我们可以使用更加先进的路由器（称为“多协议路由器”）来描述、处理和发送多协议数据包。

表22-1列出了常见的网络层协议及其相关的网络操作系统或上层协议以及其他协议，这些协议实际上是部分路由器使用的链路控制协议。总之，网桥是用来处理数据链路层协议的，那些具有和网桥相同功能的路由器称为“桥接路由器”或“桥式路由器”。

表22-1 网络层与数据链路层协议

网络层协议	网络操作系统或协议栈名称	网络层协议	网络操作系统或协议栈名称
IPX	NetWare	OSI	Open Systems
IP	TCP/IP	其他协议	
VIP	Vines	LAT	Digital DecNet
AFP	Appletalk	SNA/SDLC	IBM SNA
XNS	3Com	Netbios	DOS-based LANs

（4）路由器配置

类似于网桥，路由器的物理外形通常包含以下两种：单机独立型；安装在槽形底盘上的模块。路由器可以用来连接本地的或远程的各个LAN分段。边界路由选择可以为远端办公室中的广域网设备提供整个广域网的全路由选择功能。边界路由选择的物理拓扑有时称为“中心点和辐射状拓扑”，这是因为拓扑中的每个远端分支通过单个WAN链路连接到了一个中心点。如果某个节点的业务需要冗余链路，那么在该拓扑结构中就需要一个全能路由器，而不是边界路由器。

全能路由器位于每个HUB或中心节点处，而边界路由器或分支路由器位于每个远端节点或分支节点处。由于只连接到WAN链路，因此这些边界路由器在检查每一组数据时就只需要做出一个决定，即“如果数据的目的地址不是指向本地的，那么该数据就可以通过”。这种“非本地则通过”逻辑说明了，边界路由器实际上就是一个网桥。

4. 网关

再回到OSI模型中，我们可以发现，转发器是一种物理层（第1层）设备，网桥是一种数据链路层（第2层）设备，而路由器则是网络层（第3层）设备；这样，我们可以发现网关就是为会话层、表示层和应用层（第5~7层）之间提供互操作性的设备。转发器、网桥和路由器可以在两个LAN之间提供越来越先进的连接，而网关则在两个完全不同的计算环境之间提供透明连接。专用网关还可以在不同数据库管理系统之间进行协议转换，因此称为“数据库网关”；或者在不同的电子邮件系统之间进行协议转换，因此称为“电子邮件网关”。

网关通常是一台计算机，该计算机物理上连接了两种不同的计算环境。另外，网关上运行的是专门编写的软件，该软件可以转换两种计算环境之间的消息格式。与其他互联组网设备注重处理目的地址和尽量可能高效率发送消息不同，网关更加注重于协议的转换。

22.5 广域网

应用

1. 广域网的结构

为了更好地理解当前和正在涌现出的广域组网技术和业务，图22-11给出了一个广域网结构的简单模型，该模型定义了广域网结构的各种主要分段及其相互关系。用户需求是当前和正在涌现出的广域“网络业务”发展的主要推动力，网络业务的主要对象是企业和驻地用户。各个公司提供这些业务的主要目的是为了通过实现底层结构来创造更多的利润，该底层结构可以以用户期望的最低成本为其提供广域组网业务。

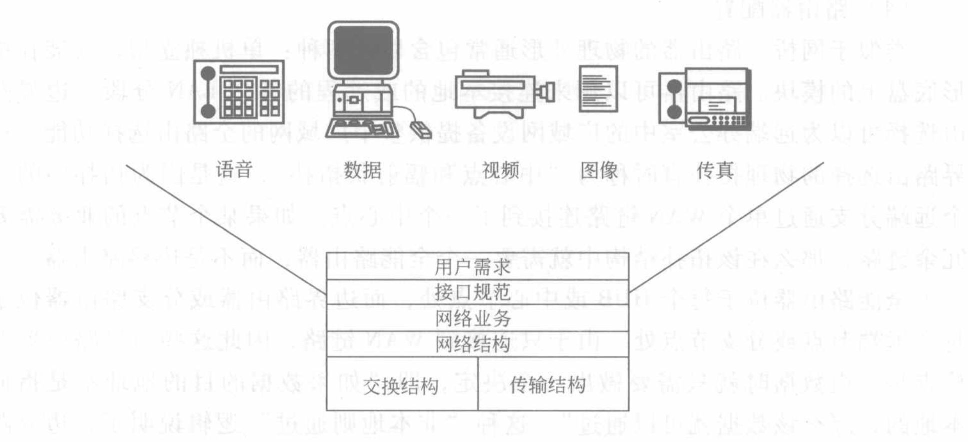


图22-11 广域网结构的主要部分框架

用户需要的是简单、透明地接入到足够宽的有效带宽上。另外，该广域网接入必须可以提供数据、视频、图像、传真和语音的传输。推动广域网业务性能和容量发展的主要动力之一就是“LAN 互联”。

2. 电路交换-分组交换

在广域网中必须有一种交换机制，否则那是不可想象的。具体来说就是，如果没有交换机制或结构，各种可能的数据源都必须直接连接到各种可能的数据目的，这是无法想象的。

(1) 电路交换

交换过程允许在消息源和消息目的地（在数据通信中有时称为“宿点”）之间临时建立、保持和终止连接。在人们熟悉的语音电话网中，每次呼叫就是通过一个中心设备（称为“交换机”）来进行路由选择的，交换机在呼叫电话和被呼叫电话之间建立了一个临时的电路。这种连接或临时电路只在呼叫持续的时间内存在，通话一旦结束，该连接也同时断开。这种交换技术称为“电路交换”，它是两种主要交换技术中的一种，用来将消息从某一端传输到另一端。在一个电路交换网中，交换专用电路负责连接两方或多方用户，同时消除了类似封装技术提出的各种源地址和目的地址信息需求。

电路交换网中的交换专用电路对于用户来说就如同是在呼叫方和被呼叫方之间直接连接了一条线路。这种临时连接所占用的物理资源在该临时连接一结束时就会被释放。如果系统用户数已经达到了可提供的最大有效用户数，那么多余的用户就听不到拨号音了。

(2) 分组交换

另一种用来将消息从某一端传输到另一端的主要交换技术就是“分组交换”。分组交换与电路交换有很多关键性的区别。首先，在分组交换网（有时也称为“公共数据网”）中，从消息源到消息目的地址之间，每次只能传输一个消息分组。分组交换网在网络结构图中是通过一个云朵图符号来表述的，如图22-12所示。云朵图非常适合用来表示分组交换网，因为我们知道数据分组（也称为“数据包”）是从公共数据网（Public Data Networks，PDN）的一端输入，而从另一端输出的。每个数据分组采用的物理路径都可能各不相同，这一点对于终端用户来说是透明的。分组交换网中的云朵图表示的是大量的分组交换机，这些分组交换机负责在源地址和目的地址之间转移数据。

数据分组是一群特别构造的数据，除了包含数据本身之外，还包含了控制和地址信息。在进入分组交换网之前，这些数据必须进行组合（控制和地址信息加上数据）；因此，在数据传输到目的地之前必须分解。这个数据组合和分解过程是由一个称为“分组组合器/拆分器（Packet Assembler Disassembler，PAD）”的设备来完成的。PAD 可以是一个独立的设备，也可以集成到 Modem 或多路复

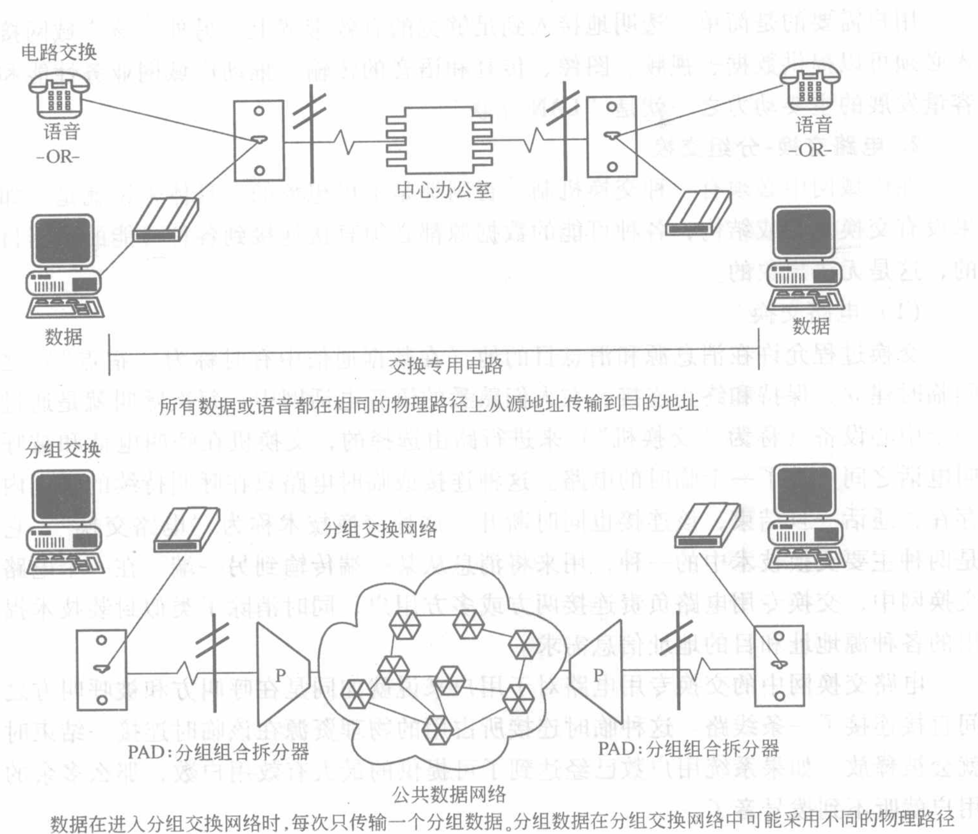


图 22-12 电路交换-分组交换

用器中。

这些 PAD 可能位于终端里面，也可能位于分组交换网的入口处。图 22-12描述了 PAD 位于分组交换网入口处的情形。在这种情形中，终端用户利用常规 Modem 来拨号增值网（Value Added Network，VAN）或在线信息业务，该 VAN 中的 PAD 在数据分组进入分组交换网传输之前可以对其进行合适的组合。

(3) 分组交换网

图 22-12 中 PDN 云朵图描述的交换机通常称为“数据交换机（Data Switching Exchanges，DSE）”或“分组交换机（Packet Switching Exchanges，PSE）”。DSE 是 Modem 和拨号传输中的数据通信设备（Data Communications Equipment，DCE）和数据终端设备（Data Terminal Equipment，DTE）类分组交换等价设备。

分组交换与电路交换的另一个主要区别就是随着分组交换网上数据传输需求

的增加，多余的用户不会被禁止接入到分组数据网中。负载过重可能会导致网络崩溃，或者产生错误和重发，或者产生数据分组丢失，但是这种现象是所有用户一起经历的；这是因为在分组交换网中，每个传输机会只能传输一个数据分组，这种机会对于所有用户是均等的；而在电路交换网中，如果出现负载过重，用户就会等待一个有效的交换专用路径。

对于任何处理分组数据的分组交换来说，每个数据分组中都包含了分组地址信息。每个分组交换通过路由选择和发送方案来读取并处理分组数据，这些路由选择和发送方案是根据数据分组的目的地址和当前网络的状态来决定的。可以用来惟一区别每个数据分组目的地的目的地址称为“全局地址”。

由于每个消息都会被分组组合器分割成多个分段，因此这些消息分段可能不会依次顺序到达消息目的地址，这是因为分组交换网中各个消息分段的传输路径速度和状态都各不相同。因此，在数据消息到达目的地址之前必须由目的地的PAD来对分段进行顺序重组。这种包含了完整源地址信息和目的地址信息外加一个消息分段信息的“自给自足”分组数据称为“数据报”。

(4) 电路交换和分组交换的业务前景对比

如果采用自顶向下模型来分析交换技术，那么电路交换和分组交换最适合放置在网络或技术层中。为了选择合适的交换技术，必须对自顶向下模型中网络层以上的所有各层（一起称为“数据层”）进行彻底的检查。这样，数据层中的关键问题就变成了：等待传输的数据的性质是什么，支持这种数据的最佳交换技术是什么？第一个与数据相关的检查标准就是“数据源”。产生这些数据的应用程序（应用层）的性质是什么？这些应用程序是面向处理的程序或批量更新程序，还是面向文件的程序？

面向处理的程序（产生的数据有时称为“交互式数据”）的主要特点就是有很短的数据突发，数据突发之后就是可变长度的暂停（由于用户读取屏幕提示符导致的）或处理过程之间的暂停。这种“面向处理的突发数据流”（最典型的例子就是银行自动提款机上的即时业务）必须由网络尽快地、可靠地进行传输。除了数据突发外，时间和可靠性上的约束条件也是决定交换技术选择的重要因素。

面向大型文件传输应用程序或批处理的应用程序与面向处理应用程序的数据特性是不同的。典型的例子就是，从区域分部到公司总部或者从本地子公司到区域分部的数据必须持续更新。这类应用程序中的数据不会产生突发，这些数据通常都是大规模的，而且都是稳定传输的。这种传输非常重要，但它不是紧急的。如果传输出现失败，检错和纠错协议会重发出现错误的数据，或者在故障点重新启动一次传输。

名词解释

带有冲突检测的载波侦听多路访问（CSMA/CD）：网络通信的一种方案。

客户：网络及其资源的终端用户，如工作站或个人计算机。

以太网：一种遵守 IEEE 802.3 标准的网络结构；基于 CSMA/CD 结构；按照惯例一般位于总线配置中，但目前更多位于基于 HUB 的星形物理拓扑结构中。

光纤分布式数据接口（FDDI）：一种组网方案，采用独立的环形结构，数据在两个环上以相反的方向传输，以实现高速传输和操作冗余。

网关：一种网络设备，用来在完全不同的计算环境之间提供透明连接。

HUB：星形物理拓扑结构中的中心，也称为“集线器”、“转发器”或“多站接入单元（MAU）”。

网络接口卡（Network Interface Card，NIC）：用来连接网络与本地工作站或设备的物理设备或电路。

网络管理：监控和分析网络流量并解决网络相关问题的行为总称。

开放式系统互联（OSI）模型：一种框架结构，使用国际标准组织开发的组网技术。

路由器：用来读取特定网络层协议从而提高网络过滤率和发送率的设备。

服务器：网络中的一种组成单元，用来促进、管理客户设备和工作站之间的资源共享。

令牌环：一种遵守 IEEE 802.5 标准的网络结构；使用星形物理拓扑结构、连续消息发送和令牌通过接入技术。

无线 LAN：一种正在涌现的组网系统，利用无线射频或红外媒质作为工作站之间的连接方式。

参考文献

[1] Bachus，K. and Longsworth，E. 1993. Road nodes. *Corporate Computing* 2（3）：54-61.

[2] Bradner，S. and Greenfield，D. 1993. Routers：Building the highway. *PC Magazine* 12（6）：221-270.

[3] Derfler，F. 1993. Ethernet adapters：Fast and efficient. *PC Magazine* 12（3）：191.

[4] Derfler，F. 1993. Making the WAN connection：Linking LANs. *PC Magazine* 12（5）：183-206.

[5] Derfler，F. 1993. Networking printing：Sharing the wealth. *PC Magazine* 12（2）：249.

[6] Derfler，F. 1993. To catch a thief. *PC Magazine* 12（16）：NE1-NE9.

[7] Derfler，F. 1994. Extend your reach. *PC Magazine* 13（14）：315-351.

[8] Derfler，F. 1994. Peer-to-peer LANs：Peer pressure. *PC Magazine* 13（8）：237-274.

[9] Donovan，W. 1993. A pain-free approach to SNA internetworking. *Data Communications* 22（16）：99.

[10] Gasparro，D. 1994. Putting wireless to work. *Data Communications* 23（5）：57-58.

[11] Goldman，J. 1995. *Applied Data Communications：A Business Oriented Approach.* Wiley，New York.

[12] Greenfield, D. 1993. To protect and serve. *PC Magazine* 12 (9): 179.

[13] Gunnerson, G. 1993. Network operating systems: Playing the odds. *PC Magazine* 12 (18): 285-333.

[14] Harvey, D. and Santalesa, R. 1994. Wireless gets real. *Byte* 19 (5): 90.

[15] Held, G. 1993. *Internetworking LANs and WANs*. Wiley, New York.

[16] Held, G. 1994. *The Complete Modem Reference*. Wiley, New York.

[17] Held, G. 1994. *Ethernet Networks: Design Implementation, Operation and Management*. Wiley, New York.

[18] Held, G. 1994. *Local Area Network Performance Issue and Answers*. Wiley, New York.

[19] Heywood, D. et al. 1992. *LAN Connectivity*. New Riders, Carmel, IN.

[20] Jander, M. and Johnson, J. 1993. Managing high speed WANs: Just wait. *Data Communications* 22 (7): 83.

[21] Johnson, J. 1993. LAN modems: The missing link for remote connectivity. *Data Communications* 22 (4): 101.

[22] Johnson, J. 1994. Wireless data: Welcome to the enterprise. *Data Communications* 23 (5): 42-55.

[23] Karney, J. 1993. Network lasers: Built for speed. *PC Magazine* 12 (20): 199.

[24] Madron, T. 1993. *Peer-to-peer LANs: Networking Two to Ten PCs*. Wiley, New York.

[25] Mandeville, R. 1994. Ethernet switches evaluated. *Data Communications* 23 (4): 66-78.

[26] Mathias, C. 1994. New LAN gear naps unseen desktop chains. *Data Communications* 23 (5): 75-80.

[27] Peterson, M. 1993. Network backup evolves. *PC Magazine* 12 (16): 277-311.

[28] Pompili, T. 1994. The high speed relay. *PC Magazine* 13 (1): NE1-NE12.

[29] Quiat, B. 1994. V. FAST, ISDN, or switched 56 K. *Network Computing* 5 (3): 70.

[30] Raskin, R. 1993. Antivirus software: Keeping up your guard. *PC Magazine* 12 (5): 209-264.

[31] Rosen, B. and Fromme, B. 1993. Toppling the SNA internetworking language barrier. *Data Communications* 22 (9): 79.

[32] Saunders, S. 1993. Choosing high speed LANs: Too many technologies, too little time? *Data Communications* 22 (13): 58-70.

[33] Saunders, S. 1994. Building a better token ring network. *Data Communications* 23 (7): 75.

[34] Saunders, S. 1994. Full duplex ethernet: More niche than necessity? *Data Communications* 23 (4): 87-92.

[35] Schlar, S. 1990. *Inside X. 25: A Manager's Guide*. McGraw-Hill, New York.

[36] Shimada, K. 1994. Fast talk about fast ethernet. *Data Communications* 23 (5): 21-22.

[37] Stallings, W. 1992. *ISDN and Broadband ISDN*, 2nd ed. Macmillan, New York.

[38] Stevenson, T. 1993. Best of a new breed: Groupware—Are we ready? *PC Magazine* 12 (11): 267-297.

[39] Tabibian, O. R. 1994. Remote access: It all comes down to management. *PC Magazine* 13 (14): NE1-NE22.

[40] Thomas, R. 1994. PPP starts to deliver on interoperability promises. *Data Communications* 23 (6): 83.

[41] Tolly, K. 1993. Can routers be trusted with critical data? *Data Communications* 22 (7): 58.

[42] Tolly, K. 1993. Checking out channel-attached gateways. *Data Communications* 22 (8): 75.

[43] Tolly, K. 1993. Token ring adapters: Evaluated for the enterprise. *Data Communications* 22 (3): 73.

[44] Tolly, K. 1994. FDDI adapters: A sure cure for the bandwidth blues. *Data Communications* 23 (10): 60.

[45] Tolly, K. 1994. How accurate is your LAN analyzer? *Data Communications* 23 (2): 42.

[46] Tolly, K. 1994. The new branch-office routers. *Data Communications* 23 (11): 58.

[47] Tolly, K. 1994. Testing dial up routers: Close, but no cigar. *Data Communications* 23 (9): 69.

[48] Tolly, K. 1994. Testing remote ethernet bridges. *Data Communications* 23 (8): 81.

[49] Tolly, K. 1994. Testing remote token ring bridges. *Data Communications* 23 (6): 93.

[50] Tolly, K. 1994. Testing UNIX-to-SNA gateways. *Data Communications* 23 (7): 93.

[51] Tolly, K. 1994. Wireless internetworking. *Data Communications* 22 (17): 60.

备注

下面两本书是关于数据通信和组网的优秀参考资料：

[1] Newton, H., *Newton's Telecom Dictionary*, Telecom Library, New York.

[2] Goldman, J. E., *Applied Data Communications: A Business Oriented Approach*, Wiley, New York.

第 23 章　印刷技术及系统

John D. Meyer

23.1　引言

印刷质量的基本参数包括：分辨率、精度、灰度和墨点的微观结构。每个实际的印刷设备在印刷过程中都具有各自内在的差异性，因此可能会由于各种干扰而产生视觉上的假象，具体表现为本底散射、墨点位置误差、空白（如由于喷墨打印机中的喷嘴故障而导致的）以及图像上的条状痕迹。印刷质量各个方面的好坏是根据人的视觉系统来衡量的，因此，印刷质量的基本参数在设计时必须根据人的视觉系统来决定哪些地方需要改进。

1. 分辨率与精度

人们有时会将分辨率与精度这两个概念弄混淆。分辨率是评价印刷质量时最常用的衡量标准，用来表示印刷设备对细节的描述能力；这种定义是非常简单的，如果具体考虑到印刷中的细节，那么分辨率的定义就会变得非常复杂了。这些细节包括：印刷线宽度的优良性、空白处与着墨处之间的过渡、曲线边沿的平滑度或者任意角度弧线的平滑度。在最简单的定义中，印刷设备的分辨率是指墨点之间的间隔，即当墨点完全覆盖时，白纸上是看不到白点的。对于方块格子中的圆形墨点，分辨率可以通过计算圆形墨点直径的二次平方根并取其倒数得到。例如，一个理想的分辨率为每英寸 300 个墨点（dots per inch，dpi）的打印机，可以以 85μm 的间隔打印出 300 个直径为 120μm 的墨点。考虑到墨点的位置误差，墨点的尺寸会比标准稍微大一些。这种定义在印刷质量最好的直线中最容易理解。分辨率为 300dpi 的直线会呈现出非常明显的波纹状边沿，尤其是以敏感的角度印刷时；这种现象对于曲线的印刷来说也是一样的。另外，宽度不断增加的直线段中会出现不连续性，因为这些直线是由整数倍的基本线条构成的，其中，线条之间的间隔为 85μm。上面的这些要点对于文字的印刷质量来说非常重要，因为文字的印刷质量取决于曲线和直线以不同宽度进行印刷时的印刷能力。

上面定义的是与人体视觉系统相关的分辨率规范；分辨率是由相同宽度的交替黑白线条之间的间隔大小决定的，这些交替的黑白线称为“线耦”；对于分辨率为 300dpi 的打印机来说，可以在每英寸的白纸上打印出 150 对线耦；这个数字严格来说不是很准确，因为由于墨点是圆形的，所以黑线条比白线条要宽一

些。由于人体视觉系统对每英寸距离上300条线条的分辨率几乎为零，因此，通过提高分辨率来提高文字印刷质量的想法是没有意义的。如果想进一步提高印刷质量，就必须考虑印刷干扰和灰度阶数等因素了。

为了集中介绍分辨率，我们先忽略对印刷材料组成成分的特殊要求，确切地说就是文字、线条以及图像和区域填充要求。如果印刷设备可以将墨点之间的间隔打印得比基本分辨率还近，那么文字的印刷质量可能会下降；这样将会造成墨点之间相互重叠，但线条宽度就会变得更加连续了。另外，在曲线的边沿处，对每个墨点中的像素调整可以消除边沿锯齿，从而提高曲线的平滑性。这样，印刷设备的最终墨点间隔就称为“精度”。例如，分辨率为300×600dpi标准的印刷设备，就可以实现300dpi的横向分辨率和600dpi的纵向精度。

2. 灰阶控制

调整页面上着墨密度的印刷技术称为“灰阶控制”。灰阶控制的方法主要有3种：改变墨点大小、改变墨点着墨密度和数字中间色技术。前两种方法的使用取决于印刷技术的本质特性，而数字中间色技术可以用于任何印刷设备。如果一个印刷设备的着墨密度可以从纯白连续变化到最高的色彩强度，那么我们就说这种设备具有“连续调色能力”。其他技术如果只能产生一定程度的色彩强度，那么它将采用数字中间色技术来产生连续调色的效果。灰阶控制的实现方法在图像印刷中非常重要，尤其是在彩色图像印刷中。最近几年，数字中间色技术已经发展到了二进制印刷技术（即单个墨点大小，无密度调节）；这样，图像的质量就更倾向于由分辨率来决定了。但是，由这些调色原理实现的印刷质量如果产生一点点的灰阶中断，那将会产生很严重的影响。

灰阶控制条件中的重要参数就是动态范围，或者称为“形象艺术范围”。该参数是根据光学密度来衡量的，其值等于反射系数的负对数。光学密度1.0表示了每个瞬间光学通量10%的反射光；光学密度2.0则表示了1%的反射光，依此类推。对于印刷材料来说，印刷面的平滑度直接影响着光学密度可以达到的最大值。如果印刷面非常平滑，如同镜面一样，那么印刷效果将会出现光泽，而且光学密度值可以达到2.4。由于平滑面会以镜面的方式进行光反射，因此，基本不会产生散射，而且色彩强度也会很饱和。这种现象在照相纸上非常明显，因为照相纸具有很好的抛光效果。如果印刷技术的光学密度范围很广，那么我们可以说该印刷技术具有很大的印刷动态范围，而且可以印刷出高质量的图像。但是，不是所有的印刷纸都具有抛光效果。例如，办公室里面用来复印的纸张就会产生漫反射。对于大多数没有加膜的纸张来说，无光泽纸张占到了3%～4%的比例，而且其最大光学密度值会被限制在1.4左右。采用这种纸张作为底板的图像质量就直接取决于着色剂在纸面上的调色技术。因此，一个印刷设备的图像质量与很多因素有关，其中包括：分辨率的设计方法、精度、灰阶控制方法、数字中间色

原理、纸张质量、着色剂以及调色技术。总之，对于图像来说，分辨率不是决定印刷质量的惟一因素。

3. 墨点微结构

印刷技术中墨点的微观性质也会对最终的印刷质量产生影响；其中，最重要的参数就是墨点的边沿斜率，称为“常规边沿轮廓”。常规边沿轮廓描述了空白处到最大着墨密度之间的过渡特性（即墨点边沿处的光学密度斜率），而且衡量了从空白处到所有光学密度之间的过渡陡峭性。某些技术（如电子照相）可以通过调整成像过程和制作过程中的各种参数来修改这些边沿轮廓。对于喷墨印刷来说，不同的纸张将会产生不同的常规边沿轮廓效果。如果边沿轮廓非常陡峭（即空白处到光学密度之间的过渡间隔非常小，如 5μm），那么该墨点就可以认为是一个很深硬的墨点，或者说具有很锋利的边沿。这种陡峭的边沿在印刷线条和文字时非常理想，因为白与黑之间的明显过渡特性非常适合线条和文字。如果这种过渡是渐进的，那么墨点就可以认为是柔和的，而且可以产生一个模糊的边沿。这种模糊的边沿如果用于文字印刷将会降低印刷质量；但是，如果用于图像的印刷，由于色调和色调变化的平滑性，柔和的墨点将会带来很好的印刷质量。

4. 混合方法

从前面的讨论中，我们并不能说文字和图像在印刷质量方面的需求就是相反的。最近几年，人们已经使用内在灰阶控制性能来提高文字印刷质量。通过提高精度和灰阶数可以在很大程度上帮助消除边沿锯齿。通过在锯齿边沿处使用灰阶控制，可以在多个像素点上产生过渡效果，这样就可以使得边沿看上去比较柔和。在有些特殊字体中，细节要求的分辨率比印刷设备本身的分辨率还要高；这些细节问题可以结合灰阶控制和常规像素控制来解决。前面这些方法在实现时需要一组非常复杂的规则，这一组规则在数据流送入印刷设备进行分级标识之前使用。这一组规则的重点是图像处理技术和人体视觉系统，而且这些规则通常都会形成一项专利。如果使用熟练，这些规则可以大大提高文字和线条印刷的质量。这些技术都拥有各种各样的商标名称，因为商标名称可以用来宣传印刷质量的先进性。

图像处理技术的应用（用来处理电子印刷技术中的内在特性）使得分辨率不再是衡量印刷质量的惟一标准了。因此，我们需要一个更加综合的衡量标准来简化印刷技术之间的区别，以便为最终的目标（即印刷质量）服务。在这样的衡量标准产生之前，工业标准测试表（单独描述印刷机的特性）中描述和实现的权衡分析过程将会提供一个预先的印刷质量衡量标准。这样的测试表还必须包含测试图像，测试图像是根据制造商提供的专利主观图像增强算法来形成的。

23.2 印刷技术

任何印刷技术都包含4个基本要素：定址、印刷物质及其存储与释放、印刷物质的转印、调色。定址是指电子数据与标印单元之间通过电子技术或光学技术进行的通信；印刷物质包含着色剂、转印/载体材料、着色剂与纸张之间的粘合剂、防止褪色的稳定剂、专用添加剂（如墨汁的抗微生物剂）。印刷物质着墨到纸张的转印过程是印刷过程中的基本原理，在这个过程中，印刷物质根据一定的规律从容积中转印到印刷纸上面。调色过程具体描述了印刷材料在纸张上的粘附、烘干或凝固并形成持久图像的过程。上述这些基本过程构成了各种印刷技术自身独有的特性。如今，常见的印刷技术主要包括两种类型：击打式和非击打式印刷技术。击打式技术是在印刷单元中通过直接的机械撞击方式来实现转印的，这个印刷单元可以是一条细丝，也可以是着色剂上的全形字符，该着色剂上有一条与纸张相接触的色带。击打式印刷技术最常见的应用就是打字机。非击打式技术覆盖的范围很广，它是通过各种方式来实现转印的，这些方式可能是接触式的，也可能是非接触式的。

23.3 非击打式印刷技术

1. 喷墨技术

喷墨印刷中的转印过程是指将墨汁中的墨滴从容积中提取出来，并给定一个具有足够精度和量值的速度，将墨滴驱动到非常接近但不接触打印头的一个衬底上。喷墨技术主要包含3种：连续喷墨、静电喷墨和按需喷墨（Drop On Demand，DOD）。连续喷墨印刷技术主要用于商业系统中，因为它具有很高的喷墨速度；静电喷墨印刷技术目前还没有得到广泛应用，但已经用于传真记录系统中了；按需喷墨印刷技术已经广泛用于办公室和家庭打印中了，因为这种技术非常简单，而且容易实现彩色打印。

2. 连续喷墨技术

连续喷墨技术利用了墨汁不稳定性产生的自然分散过程原理，当墨汁液体在一定压力下通过一个小孔时，就可以形成连续的喷墨过程了。这种连续喷墨的原理来自于墨滴的表面张力和粘性之间的相互影响，而且是以拟随机的方式实现的，除非使用外部激励。Rayleigh 首先对自然分散过程进行了研究，他是通过墨汁不稳定性的增长速度来描述该过程的。其中，墨盒直径为 D，速度为 V，外部激励的频率为 F。Rayleigh 指出，墨汁不稳定性的最大增长速度对应的频率为：$F = V/4.5D$。通过在 F 频率下激励墨盒，就可以得到一条整齐的墨滴流。目前，

提供这种激励的典型方法就是利用压电式换能器作为一个完整的打印头部件。

为了在印刷过程中有效使用这些墨滴流，必须在“中断面”处对其进行充电；这个过程是通过在接近墨盒的分散区域放置两个电极来实现的。放置好电极之后，在喷墨方向的下游处设置偏转电压来引导墨滴到达基底的方向，或者进入一个墨滴收集装置以供循环使用。最早的印刷技术就是利用这种对墨滴进行充电的原理，并利用偏转电压来引导墨滴到达纸张方位，从而沿着某个方向打印出具有标准高度的字符（见图23-1）。最新的印刷技术主要集中在产生带电的墨滴流，并利用印刷电极（高压）来偏转多余的墨滴，从而进行回收以供循环使用（见图23-2）。这种技术称为“二进制带电连续喷墨”技术，有助于构建各种喷嘴排列，而且现在已经有了很多页面宽度喷嘴排列的实现方式。

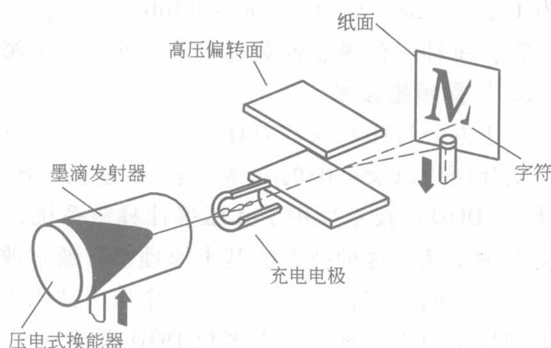


图23-1 利用连续喷墨技术打印字符

注：偏转板使用模拟电压来将墨滴引导到理想的位置上，多余的墨滴不会被偏转，从而进入一个收集器中，以供循环使用

（来源：Durbeck，R. C. and Sherr，S. 1988. *Hardcopy Output Devices*. Academic Press，San Diego，CA. 引用已经过允许。）

二进制带电连续喷墨技术由于具有很高的墨滴速度，因此，可以形成简单的灰阶控制技术。纸面上墨点的大小可以通过在相同的位置打印1个墨滴或N个墨滴的方式来进行调整，其中N是所需不同墨点大小的数量。通过高频和微型墨滴，我们可以得到足够的灰度阶数，这样就可以实现一定速度的全灰阶控制印刷。目前，这种方法在精度为200~300像素/英寸的条件下，最高可以提供512个灰度阶数。为了使墨滴的尺寸尽量小，我们采用直

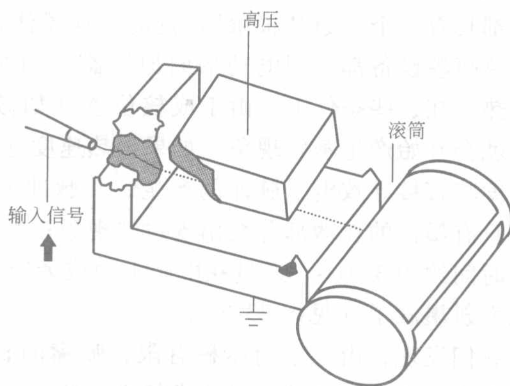


图23-2 二进制带电连续喷墨

注：墨滴在“中断面”处充好电之后，被引导到旋转定影辊的纸面上。多余的墨滴将通过高压电极转向进入一个收集器中

（来源：Durbeck，R. C. and Sherr，S. 1988. *Hardcopy Output Devices*. Academic Press，San Diego，CA. 引用已经过允许。）

径为10μm量级、压力为500~700lb/in^{2}的玻璃毛细管。彩色打印头上包含4个毛细管，每种颜色墨盒各对应一个，外加一个黑色。

3. 按需喷墨技术

对于办公打印和家庭打印来说，连续喷墨技术过于复杂（如启动和关闭流程、墨汁回收以及有限的喷嘴数量），这就直接促进了按需喷墨（DOD）技术的发展。（DOD）技术采用了无压墨汁释放系统，顾名思义，该系统只有在需要时才会提供墨滴。这种技术的基本原理是，通过改变墨汁供给渠道的容积或者喷嘴附近墨室空间的容积，从而产生一个压力波来发射墨滴。而墨汁的回填补充是通过毛细作用力来实现的，大多数DOD系统在工作时都会在墨盒上施加轻微的负压。压力波的产生机制决定了这种印刷设备的设计过程；目前，在常见的DOD系统中，压力波的产生机制主要包括两种：一种是利用加热墨汁升华作用产生的压力脉冲；另一种是采用压电材料，这种材料在电压的作用下会产生变形。

利用热流蒸汽来产生压力脉冲的设备包括热喷墨打印机或泡沫打印机，这些设备的名称是由生产商确定的。由于按需喷墨技术依赖毛细作用来进行墨汁补充，因此其工作效率就比连续喷墨打印设备低很多。这一点就影响了多个喷嘴打印头的打印速度，从而降低了传动系统的连贯性。如果想避免出现印刷重影，各个喷嘴都必须进行严格校准。

4. 热喷墨/泡沫DOD打印机

当墨汁在极端的速度（如500×10^6 W/cm^2）下被加热时，墨汁将会进入一个短暂的亚稳态；在亚稳态下，常压下的墨汁温度可以远远超过沸点。这个高温与沸点之间的温度差称为“过热度”。这个高温过程不会持续很久，因为所有的液体都具有一个“过热极限”。这时，在墨汁容积内就会产生成核现象和蒸发现象。热喷墨设备都采用电动平面加热器（如面积为50~60μm^2）在墨汁表面进行加热；在这些条件下，由于成核位置（如微细粗糙面）的存在，在加热器的表面就会开始产生蒸发现象。如果加热速度选择得当，这个过程就非常可靠。加热速度将直接导致电子脉冲的产生，其脉冲宽度为3~5μs。在这个脉冲时间段内，只有墨汁的亚微部分会出现过热现象；其结果就是蒸汽脉冲超过气压，而且持续时间约为3/4μs。通过在电阻上面或旁边直接放置一个喷嘴，该压力脉冲就可以发射墨滴了（见图23-3）。

我们发现，由于墨滴体积有限，喷嘴的直径和平面电阻具有类似的线性尺寸。因此，传动装置就可以非常精密，这样，就可以实现多个喷嘴印刷头了。电阻是采用照相印刷技术来制造的，这种技术在IC产业中很常见，而且电阻衬底是硅质的，其上有一层很薄的二氧化硅绝缘层。这样，各个喷嘴之间的精密排列就可以得到保证，而且在打印头上还可以通过专用集成电路来实现更复杂的功能；对于扫描打印头来说，这是一个很有价值的应用方案，因为扫描打印头采用

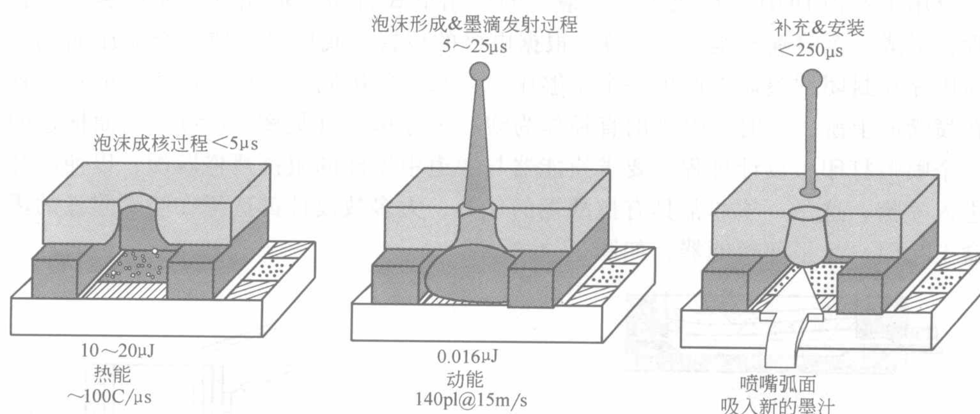


图 23-3　热喷墨技术的墨滴发射过程

注：各个时段和各种参数值用来表示以下 3 个过程：成核过程、泡沫形成过程、喷射形成过程，其后是墨滴发射以及补充过程

柔性打印电路来实现互连。以上这些特征使得在单个彩色印刷机中，可以实现 300 个打印头或更多的喷嘴。这种技术的精密性带来的好处还有墨汁的供给渠道可以完全集成到打印头里面去；这样，当墨汁耗尽时，用户就可以直接替换打印头，以免除维护操作。由于在替换打印头的过程中，纸张上的细微颗粒会进入喷嘴而且偶尔会驻留在那里，因此打印头的替换会为用户增加一定的成本。这时我们可以使用半永久性的打印头，该打印头使用可替换墨盒的墨汁。因此，热喷墨技术的设计必须考虑很多因素，如：维护的频率、操作成本、打印机的使用频繁度、打印材料的类型等等。在介绍 DOD 技术的章节最后，我们将会对这类打印机墨汁和纸张中最重要的内容进行讨论。

5. 压电式 DOD 打印机

如果对晶体进行机械拉伸时，晶体结构将会产生自发偶极矩，同时晶体结构会发生形变；那么，这种晶体结构就称为“压电结构”。通过在合适的晶面上施加电压也可以反过来使压电材料产生形变。压电陶瓷在制造过程中将会确立一个偏振方向，这样各种应用就可以利用这个内部偏振方向来产生机械位移。根据应用中的方向，压电材料可以进行横向或纵向扩展和压缩；因此，这些压电材料可以作为换能器而且已经广

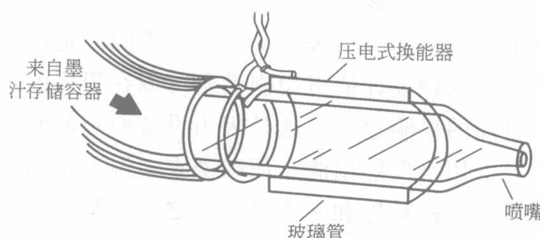


图 23-4　套管压电式喷墨技术

注：早期用于发射墨滴的压电式换能器模型

（来源：Advanced Technology Resource Corporation.）

泛应用于各种 DOD 打印机中了。最早的应用形式就是在玻璃管外面包裹一个套管，形成一个喷嘴（见图 23-4）。根据电极的位置，我们可以使用径向压缩或横向压缩在封闭的墨盒中产生一个足够压力波来发射墨滴；这种方法将换能器放置在喷嘴的上游，同时将喷嘴的直径作为线性尺寸单位（见图 23-5）。上面描述的多个喷嘴打印头设计过程，要求换能器与通道中墨汁的阻抗严格匹配，以便墨汁进入喷嘴；这是一项非常具有挑战性的任务，大多数设计都选择在喷嘴附近的墨盒上粘附一个平面换能器，如图 23-6 所示。

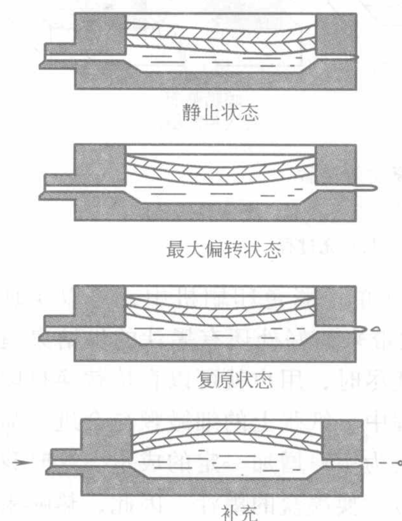


图 23-5 压电式打印头的墨滴发射过程

注：通过对毛细管上的压电晶体进行偏转来发射墨滴。在实际应用中，由于尺寸原因，压电驱动器实际上处于喷嘴的上游位置

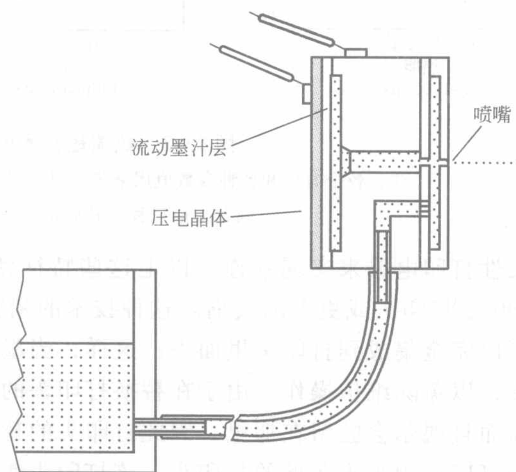


图 23-6 Stemme Larson 电动 DOD 喷墨技术的设计

注：在喷嘴处，对压力脉冲和墨盒进行直接连接

（来源：Advanced Technology Resource Corporation.）

将压电换能器通过一个墨腔直接连接到输出喷嘴的方法自诞生以来已经得到了很多改进和发展。在有些设计中，气流在喷嘴处形成了通道，这样气流在喷出时就可以将墨滴带出来，从而提高墨滴的定向稳定性，同时还可以加快墨滴的喷射速度。这种设计使得印刷设备可以在很低的换能器偏转电压下工作，而且可以产生更高的墨滴速度，这是因为设备归位的时间减少了。这样，压电设备就可以在高温下用来发射墨滴，而墨滴在常温下是固态的。固态墨汁在高温下被熔化后将形成一定的粘性和表面张力，之后将被输送到压电驱动墨腔中；最后，墨汁在喷射到纸面上之后完成凝固。

压电式打印机的最新应用就是利用压电式换能器进行纵向模式工作。其中，换能器是由单块压电陶瓷组成的，该压电陶瓷是一排一排的传动杆结构。布置恰

当的电压位置就可以推动传动杆进行纵向伸展；另外，还可以将传动杆的一端粘结一个薄膜来构成一个墨滴发射腔，这样就可以在发射腔中形成类似于前面设计中的压力脉冲（见图 23-7）。为了实现足够高的压力，可以使用隔膜，该隔膜的尺寸必须比输出孔的直径足够大。这种设计使得高密度的打印头需要多行喷嘴（见图 23-8）；因此，这种设计目前只能在液态墨盒中实现了。

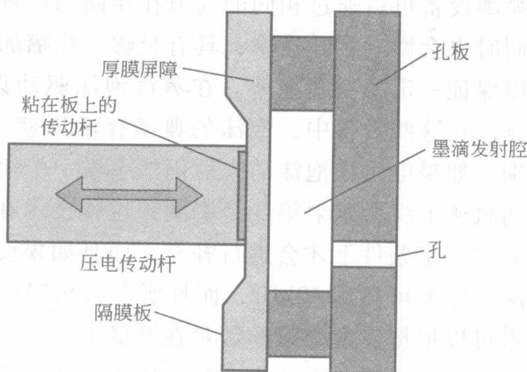


图 23-7 墨盒与传动装置示意图

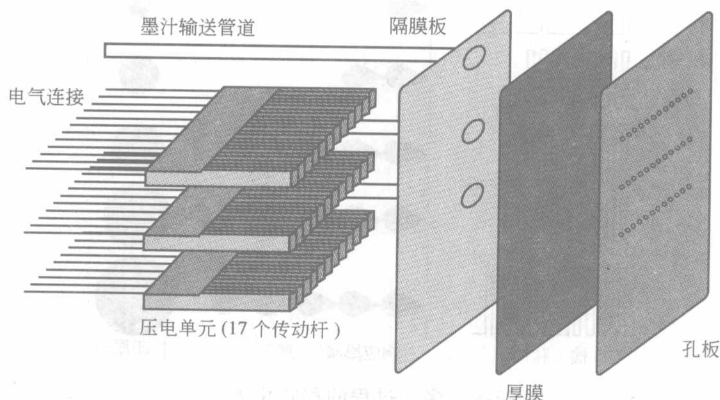


图 23-8 多墨盒设计的分解图

6. DOD 喷墨打印机的灰阶控制法

DOD 设备中的墨滴速度没有连续性，也没有压电式喷墨设备中的墨滴速度快，这就表示灰阶控制的实现策略也不尽相同。灰阶控制技术的基础是少量的灰度阶数，当结合了数字调色原理（如误差扩散、聚集的或分散的墨点、蓝色噪声抖动）后，就可以生成满意的灰度阶数。必要的灰度数量、可实现最大调节特征的相关位置（如最大墨点尺寸或密度）、数字调色技术中的专用技术等等都是目前研究非常活跃的领域。在介绍这些技术的专著中有很多专利，而且制造商一直在设法利用相关的技术来区别它们的产品。当结合了前面印刷质量章节中介绍的分辨率增强法后，具有中等分辨率（如 300dpi，每个灰度为 2bit）的打印机就可以打印出非常清晰的图像、文字和图形。

在 DOD 设备中，有很多种方法可以用来调整墨点的密度或尺寸。对于压电式设备来说，脉冲宽度调整已经证明了可以产生不同尺寸的墨滴。所有的 DOD

喷墨设备可以通过相同的通道在相同的位置根据需要的次数反复发射墨滴，但这同时也会影响吞吐速率；具有足够多喷嘴的打印头可以实现这个过程，同时还可以保证一定的吞吐速率。在蒸汽泡沫驱动设备中，只有泡沫快速吸合这一个优点。在这些设备中，泡沫的典型存活周期（从汽化到泡沫消失）处于20μm级别。如果电阻在泡沫消失后仍然会被一个短时脉冲作用，那么在第一个墨滴发射的轨迹上将会发射第二个墨滴，这种技术在相关著作中称为“多滴”。墨腔通常只在一定条件下才会进行补充，而且如果设计合理，这种方法还可以发射多个墨滴，频率可高达40kHz，而且所有墨滴都具有相同的尺寸（见图23-9）。这种技术可以根据发射墨滴的数量在基底上产生不同的墨点尺寸，这一点对于大多数压电式设备来说是无法实现的，因为其传动装置的归位速度太慢。如果没有调色原理可供参考，那么就需要使用多个喷嘴来调节不同的色彩稀释。

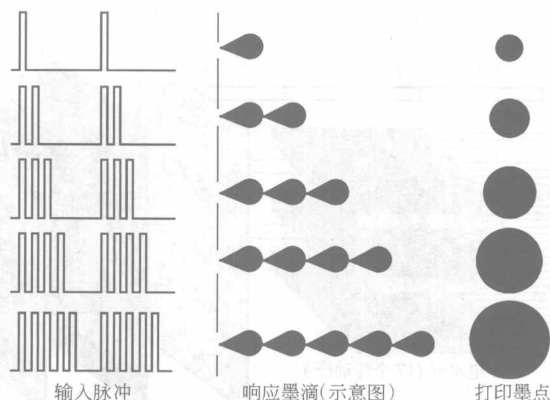


图 23-9　多滴过程的程式化表示法

注：每个输入脉冲都由一组驱动脉冲组成，这些驱动脉冲是专门为最后的墨点尺寸选定的

（来源：Durbeck, R. C. and Sherr, S. 1988. *Hardcopy Output Devices*. Academic Press, San Diego, CA. 引用经过允许。）

7. 喷墨设备的墨汁和纸张

当使用液体油墨时，纸张的特性对印刷质量的好坏将有很大的影响。墨滴被纸张吸收时，纸张的内部结构和表面张力将决定墨滴的尺寸、形状和微观结构。微观结构下的纸张是由纵横交错的纸纤维构成的，中间填充的是浆料和化学粘合剂，其本质特性是各不相同的。图23-10给出了不同体积墨滴在纸张上产生的不同墨点的大小示意图。注意，当墨滴体积较小时，墨点尺寸的曲线是非线性的；而当体积较大时，墨点尺寸的曲线接近平坦或高增益。这就表示，改变纸张的质量就可以很容易地影响印刷的质量。为了有效控制这种现象，有些纸张会在表面增加一层粘土状的薄膜材料，该材料中包含了漂白剂，通常为荧光粉。这一层薄膜材料在纸面上将会产生一个多微孔的结构，该结构比纤维素纤维结构更加

均匀。

在敷膜纸张上形成的墨点将是圆形的，而且比不敷膜纸张上的墨点更加稳定。在不敷膜的纸张上墨汁会由于毛细作用的缘故顺着纤维产生扩散，从而产生墨点“羽化”现象。在这种情况下，墨点的边沿处就会产生晕染效应，从而印刷效果也将变得模糊。这种现象在文字打印中非常严重，因为文字的边沿都是非常陡峭的。“羽化”现象在静电拷贝用的纸张中很常见；而办公用的证券纸通常因为添加了薄膜，因此很少出现“羽化”现象。

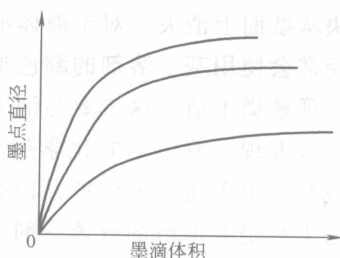


图 23-10 纸张响应曲线

注：最下面的曲线表示的是具有薄膜的纸张，其墨点扩散最小

（来源：Durbeck，R. C. and Sherr，S. 1988. *Hardcopy Output Devices*. Academic Press，San Diego，CA. 引用经过允许。）

各个制造商都采取了很多方法来尽量减小由于纸张原因而对印刷质量产生的影响。一种方法就是利用加热器，因为墨滴在喷射到纸面上之后，纸张不会立即吸收，这中间有一个过渡时间，称为“润湿时间”；润湿时间在不产生吸收时可以持续 80μs；如果立即对打印头经过的区域附近进行加热，就可以通过蒸发载体来有效“冻结墨点”。这样，打印机就不会对造纸原料的特性过于敏感，而且可以实现均匀的、高质量的印刷效果。另一种方法就是利用表面活性剂来替代化学粘合剂（该活性剂可以提高渗透速度），同时利用高气压添加剂来加快粘合剂向空气中散发的速度。因此，如果墨点的渗透速度很快而且不会向四周扩散，那么墨点的标准大小就可以得到保证了。对于按需喷墨打印机来说，我们建议全面检测造纸原料的性能。在有些情况中，我们发现只有使用制造商指定的纸张，才可以得到高质量的印刷效果。

结合前面介绍印刷质量的分析，我们必须记住纸张的选择将会影响到打印质量的动态范围。对于文字印刷来说，光学亮度必须至少为 1.3～1.4；而对于图像印刷来说，动态范围越广越好；而且如果图像质量非常重要时，就会采用比复印纸张质量还好的专用敷膜纸张。很多有效的敷膜纸张的表面仍然不是很光滑，这样就会对各种光产生反射，从而限制印刷质量的动态范围。有些制造商现在可以提供高光泽的图像印刷纸张，这种纸张具有塑料基底，并在基底上面覆盖了一层特殊的材料，这样墨汁就可以直接吸收到衬底上，留下一个高光泽的表面。这种技术大大提高了照相纸印刷的动态范围；而且这种衬底可以使喷墨打印机打印出高饱和、绚丽的色彩，而且其色彩范围非常广；如果主要用来打印图像的话，建议使用这种基底。

除了印刷质量外，对喷墨打印机还有很多其他要求。例如，喷墨打印机必须提高可靠的设备操作和耐久的图像；也就是说，图像不能很快出现褪色或者说不

能很快从纸面上消失。对于液体油墨来说，这个要求是具有一定挑战性的，因为我们经常会使用基于溶剂的颜色加亮笔来标识特定的打印文档；这些溶剂会导致油墨出现模糊不清，这主要与油墨的化学成分和基底上着色剂的覆盖方式有关。我们可以发现，着色剂的化学成分是要点，因此目前有很多关于这个问题的研究。现在有很多防褪色染料，但是很多染料都与墨汁溶剂（如水）不兼容，甚至有些染料是有毒溶剂或诱变剂。

23.4 热印刷技术

利用接触式打印头的热能控制来触发物理成像过程或化学成像过程的印刷技术可以纳入到热印刷技术的范畴中来；目前，主要有4种热印刷技术：直接加热、直接热转印、染料扩散热转印和电阻色带热转印。

1. 直接加热技术

直接加热技术是最古老、应用最多的热印刷技术。成像过程的原理是通过对热变色层进行加热来获得相应的色彩，该热变色层附着在纸面上，厚度约为10μm。这种热活性层的粘合剂中包含了分散的无色染料和酸性物质；当对这些物质进行加热时，就会产生化学反应，无色染料就会变成可以看见的有色图形。这个过程的关键在于打印头的设计，打印头可以是页面宽度的排列形状，也可以是垂直的排列形状，或者是扫描打印头。目前，主要使用两种技术：厚膜打印头和薄膜打印头。厚膜打印头的电阻材料厚度在10~70μm之间，电阻材料外面覆盖了一层厚度为10μm的玻璃，从而提高耐磨性；而薄膜打印头与热喷墨打印头非常相似，它们都采用厚度为1/10μm的电阻材料（如氮化钽）和厚度为7μm的二氧化硅耐磨层。薄膜打印头的分辨率可以高达400dpi。在任何一种情况中，电阻材料都可以循环使用，通过电加热脉冲可以实现温度的变化范围从室温（25℃）到400℃。总之，薄膜打印头在能效转换、印刷质量、响应时间和分辨率等各方面都非常出色。因此，薄膜打印头经常应用在那些需要高分辨率的系统中，而厚膜打印头经常应用在商业系统中，如条形码、飞机票、传真等。

2. 直接热转印

直接热转印技术直接将涂蜡层熔化到纸上面（见图23-11a和图23-11b）。涂蜡层中包含了着色剂，厚度为4μm，位于聚酯薄膜上面，该聚酯薄膜的厚度通常约为6μm。前面描述的热打印头会将色带挤压在纸面上；当加热单元被脉冲触发之后，涂蜡层就会被熔化并粘连在纸张上。这个过程实际上是双重的，涂蜡层转印的面积可以利用形变电阻（可以产生沙漏现象）来进行调节。通常，转印过程可以实现分辨率为300dpi的页面宽度，而垂直方向的精度可达到600dpi。在桌面打印机和便携式打印机扫描打印头的设计中，加热色带通常会封

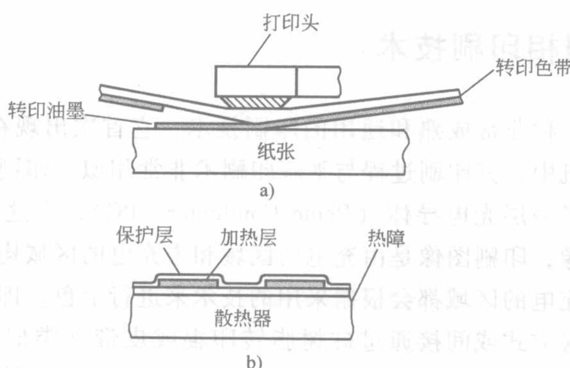


图 23-11　涂蜡层转印过程

a）打印头和色带之间紧密接触，纸张必须非常光滑　b）薄膜热打印头的设计要点，热障在加热脉冲时间内隔离了温度，但允许在脉冲之间释放热量

装在色带盒中。对于所有热打印机来说，功耗都是一个关注的重点；在直接热转印技术中主要是通过减小色带的厚度来降低功耗。

3. 染料扩散热转印技术

在染料扩散转印技术中，染料将通过升华和扩散过程从色带转印到纸张上。染料转印的量与提供的热量成正比，这就是一种连续调色技术。这种技术已经用于卤化银照片、制图和预压保护层中。所有热打印机中，能量都是通过瞬间加热过程来实现转印的；这个过程是由扩散方程式来控制的；而且根据加热脉冲的不同长度，这个过程可以在短距离内产生很大的温度梯度，或者在电阻周围的大范围区域内产生很小的温度梯度。因此，大多数设计重点都集中在了各种敷层的厚度上了，以便更好地引导加热过程。热染料升华转印过程最终将会得到比较柔和的墨点边沿；这一点非常适合图像的印刷，但不适合文字的印刷，因为加热脉冲越短，产生的墨点边沿就越陡峭。

4. 电阻色带热转印

电阻色带热转印技术类似于直接热转印技术，这一点体现在热塑性油墨是通过热量来转印到衬底上的。色带由三层组成：聚碳酸酯和碳黑电导衬底（厚度为 16μm）、铝膜层（1000～2000Å）和油墨层（厚度为 5μm）。其中，铝膜层相当于一个地回面。当电流从打印头中的电极流经聚碳酸酯/碳黑层到达铝膜层时，就会产生热量；其中，打印头与色带衬底相互接触。打印头上施加的高压可以确保与纸张紧密接触，因此纸张的平滑性要求不是非常严格。印刷好的字母可以通过以下方法来消除：即在所有电极上施加一个较弱的电压并对油墨进行加热，当油墨被加热到刚好脱离纸面时，打印头掠过就可以消除字母。这种印刷消除技术不能直接用于彩色印刷中。

23.5 电子照相印刷技术

电子照相是一种非常成熟和通用的印刷技术，它首次出现在1960年，当时应用在办公复印机中，其印刷过程与平版印刷术非常相似。印刷板是一个圆筒或皮带，其上覆盖了一层光电导体（Photo Conductor，PC），在这层光电导体上就可以形成印刷图像，印刷图像是由充电的区域和未充电的区域构成的。无论是充电的区域还是未充电的区域都会根据采用的技术来进行上色，即调色过程。图像可以直接通过接触方式或间接通过硅树脂转印辊或皮带（类似于平版印刷技术中的胶印定影辊）的方式来印刷到纸面上。早期的复印机是利用几何光学的原理来将待印材料的图像复制到光电导体上的；如果将几何光学原理替换成扫描激光束或者LED线性阵列（可以进行电子调节），就构成了今天的激光打印机。到目前为止，印刷技术已经从桌面的办公打印机技术（4~10ppm）发展到了高速的商业打印机技术（超过100ppm）；尽管商业打印机可以实现E-尺寸的印刷性能，但它在彩色和黑白印刷中的最大印刷宽度只达到8.5~17in。

1. 印刷过程

电子照相印刷包含了一系列连续的相互协同过程，如果要实现高质量的印刷效果，必须对这些过程进行统一优化。电子照相过程如图23-12所示，具体阐述如下。

1）对光电导体进行充电，得到一个均匀的静电面；这个过程可以通过很薄

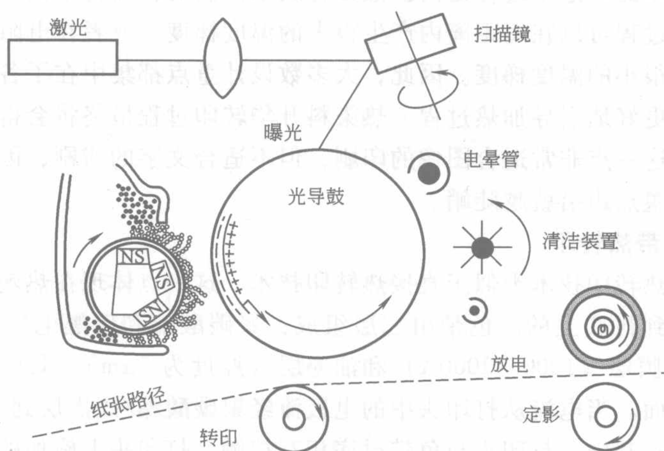


图23-12 电子照相过程示意图

注：其中对偶元件由热轧定影元件和用于充电和清洁的电晕组成

（来源：Durbeck，R. C. and Sherr，S. 1988. *Hardcopy Output Devices*. Academic Press，San Diego，CA. 引用经过允许。）

的电晕来实现，其中部分屏蔽线上的对地电压超过几千伏（电晕管）。如果是正电压，那么带正电的表面电荷来自于屏蔽线附近的电离作用；如果是负电压，那么带负电的表面电荷的产生过程就非常复杂，其中包含了二次电子发射、离子碰撞等等；该过程还会导致非均匀放电。电晕接地屏蔽罩的设计对于电荷的均匀产生非常重要。为了限制产生臭氧，很多办公打印机（<20ppm）都采用了一个电荷定影辊，该定影辊与光电导体相互接触。定影辊和光电导体之间的缝隙会产生一个局部的小型放电过程，这样可以将臭氧的数量降低两到三个数量级。

2）充好电的光电导体会被曝光从而形成图像，该图像与背景之间会有一个非常高的电压差。光电导体的特性与电子和空穴的产生直接相关，因为该曝光成像过程是依靠光和转移到光电导体表面的电子或空穴来形成图像的。这个过程实质上是一个电子照相过程，而且其转移曲线是卤化银的 H 和 D 曲线。如果想得到最佳的印刷质量，放电过程必须非常迅速，而且越彻底越好，以便在充电区域和未充电区域之间产生足够明显的电压差；暗衰减必须保持最小，而且光电导体必须可以维持重复的充放电过程，而不会出现损坏。除了必须及时适应曝光光波的波长外，光电导体的表面还必须具有抗磨损性，以便抵抗温度和湿度变化产生的影响，同时还必须将着色剂全部释放到纸面上。这样，放电区域或充电区域就都可以作为印刷过程的图像源。目前，广泛使用的（尤其是在激光打印中）图像源是放电区域。

早期的光电导体对可见光波长非常敏感，而且主要依赖于硫磺、硒和碲合金。随着二极管激光扫描仪的使用，对近红外光灵敏度的需求直接导致了有机光电导体（Organic Photo Conductor，OPC）的出现，OPC 由多层材料构成，包括亚微型厚电荷产生层、电荷转移层（厚度为 30μm）。这种结构优化了印刷过程，目前已经得到了广泛使用。OPC 使用了钝化层或耐磨层，但是由于这种钝化层过于柔软，在转印过程中会产生磨损。在很多桌面印刷设备中，光导鼓嵌入在一个可替换的色带盒中，该色带盒中装入了足够的碳粉；这种结构为用户提供了很大的便利，类似于热喷墨打印机中的可替换打印头。

3）成像过程是通过将曝光的光导面与碳粉颗粒相互接触来实现的，这些碳粉颗粒是带静电的。静电吸附作用将这些颗粒牢牢吸附在光导面上，从而形成了图像。另外，碳粉颗粒的均匀性和供给稳定性非常重要，这样才能保证与印刷过程保持一致。目前，主要采用两种成像方法：对偶元件和单成分碳粉；前者主要用于高速印刷，后者主要用于桌面打印。对耦元件方法主要采用了磁性碳粉颗粒（10μm 量级）和可磁化的圆珠载体（尺寸为 100μm）。机械摩擦搅拌可以使碳粉颗粒带上电荷；当把磁性碳粉颗粒与定影辊进行接触时，就可以形成多条线条状的颗粒链；这些从定影辊上延伸出来的密集颗粒阵列称为“磁性刷”，而且旋转时与充好电的光电导体相互接触（见图 23-10）。然后，碳粉就会被吸引到电

荷属性相反的区域，同时一个传感器控制的补充系统就会用来维持碳粉与圆珠载体数量之间的合适比例。

单成分显影原理简化了上面的过程，因为单成分显影方法中不需要圆珠载体、填充系统和附带的传感器。对于更加精密的显影系统来说，单成分显影方法主要有两种实现方式：磁化和非磁化。磁化方式中仍然会形成一个磁性刷，但是该磁性刷是由碳粉颗粒单独构成的。最常见的应用就是在定影辊的金属套管上添加一个振荡电压，碳粉刷与光电导体不接触，但碳粉颗粒产生的雾状是在颗粒沿着振荡电压电场方向进行振荡时形成的。非磁化方式就是当前应用的打印机原理，不过，如果想在理想的打印速度下得到均匀的显影，就必须提供充好电的碳粉和足够的速度，这方面的技术还存在一定的挑战性；另外，由于没有磁性添加剂，因此成本更低，透明度更高（对于彩色印刷来说）。

液态显影技术可以巧妙地回避颗粒大小的限制和某些颗粒刷技术的要求。碳粉颗粒分布在烃基载体上，并通过电偶层进行充电，该电偶层是当碳粉被放入溶液时产生的。例如，液态碳粉通过定影辊与光电导体相互接触；然后，颗粒转印机制（如电泳分离法）就会将碳粉转印到相应的图像区域。液态流动现象是这类打印机面临的主要挑战。到目前为止，液态显影技术只应用到了复杂的液态防漏商业系统中。这种技术可与平版印刷技术产生竞争，而且已经应用到了彩色校样技术中。

4）转印和定影阶段对碳粉和光电导体提出了更高的要求。碳粉在释放到纸面上时必须非常清晰，而且形成的图像必须能持久不褪色（定影）。尽管某些商业系统中可能会采用齐纳发光管的发光定影技术，但是大多数定影技术都采用热压技术。碳粉颗粒必须充分熔化并混合均匀形成一层薄膜，该薄膜将紧紧地粘附在基底上。熔化颗粒的粘性（即表面张力）和颗粒的大小都将会对转印和定影过程产生影响。这个过程中存在的设计挑战就是如何避免广泛使用热技术，并限制压力以避免平滑化，也就是避免对纸张进行研光或卷曲。加热研光定影过程利用一个加热合成塑胶定影辊与一个未加热反向定影辊之间形成的辊隙来对发黄的纸张进行研光定影，该未加热的反向定影辊上可能包含合成塑胶成分。有些设计中还在加热定影辊上使用了硅油薄膜来辅助熔化碳粉从定影辊表面上释放出来。在这种情况下，不可避免会产生一些液态流动现象，而且硅油也可能会渗透塑胶，从而影响塑胶的物理特性。因此，材料革新技术在电子照相技术中扮演了很重要的角色。

5）最后一步是指在下一次转印的充电和成像过程之前，将剩余的碳粉从光电导体上清除。常用的清除工具包括纤维刷、磁性刷和刮板。这个清除过程包括消除光电导体上的剩余电荷和清除背景板上的碳粉。清除下来的碳粉存储在一个废料容器内。清除过程中，光电导体表面的硬度对于清除效率非常重要。彩色激光打印机中，成功的清除过程非常重要，因为色彩对比度会使背景的散射十分明显，例如，一个均匀的黄色区域中的绛红色碳粉背景就会产生很明显的散射。

2. 墨点微观结构

在图像的微观结构中，碳粉原料的设计、显影技术和光电导体的特性都非常重要。因此，碳粉颗粒越小越好，而且其标称直径的分布最好是紧密分组类型。碳粉的组成成分与各种印刷过程和专利技术有关。对碳粉的基本要求是具有统一的高荷质比、高透明度（彩色）、紧密分组分布和颗粒无错误电荷属性；后面几个要求是为了避免产生影响印刷质量的背景散射。最新的碳粉生产技术可以通过电荷控制添加剂来控制上面这些参数；其中，添加剂是为了获得合适的充电电量和电荷属性。激光打印机中的灰阶控制就是通过调节二极管激光器的脉冲宽度来实现的。转印曲线的形状和陡峭性与光电导体属性、显影过程和碳粉特性有关；其中，该曲线将曝光过程与显影密度关联起来。因此，我们可以得到高斜率或低斜率的转印曲线。对于文字印刷来说，理想的斜率曲线是陡峭的；但对于图像印刷来说，理想的斜率曲线是平坦的。由于显影过程的稳定性与周围的温度和湿度有关，因此，实现一个没有印刷重影而且具有稳定灰度的彩色激光打印机是非常具有挑战性的。

23.6　磁成像和离子成像技术

这两种技术都是以碳粉为基础的，但使用了不同的着墨和转印介质。光电导体被一个很薄的可磁化介质或坚硬的绝缘层替代，如离子成像打印机中使用的阳极氧化铝。磁成像打印机可以通过改变打印头磁极间隙的磁场方向来控制打印头，该打印头在可磁化介质中可以产生磁通量转换；这种磁通量转换是强磁场梯度和磁场强度产生的原因。显影过程是通过磁性刷对磁化碳粉的控制来实现的；碳粉颗粒被磁化后，会被具有强磁场梯度的介质吸附过去。转印和定影过程与电子照相技术中的相同。离子成像打印机是通过包含绝缘电子源的打印头来对覆盖了绝缘层的定影辊进行转印的；在打印头的一个微型腔体中，可以通过射频电场中的空气击穿来产生电子；其中，电子束是通过屏蔽电极来聚焦的，微型腔体的作用如同一个真空管，使电子的运动不受空气分子的干扰。底盘的作用是由绝缘镀金定影辊控制的，其电压为接地电压。电荷图像是通过单一成分碳粉来显影的，之后是“刺穿”阶段即转印和定影操作，这个过程不会受到温度的影响。这两种技术都需要一个清除过程：对于离子成像来说，清除过程就是机械刮屑过程；对于磁成像来说，清除过程就是磁化清除过程。

23.7　系统要点

由于各种打印材料各不相同，因此，打印机中的数据传输和处理非常重要。

打印输出可能包括印刷版面样式、计算机制图以及黑白或彩色的自然图像或合成图像；由于这些信息过于复杂，因此无论是在计算机中还是在打印机中，都需要进行大量的处理。应用软件与打印机之间可以通过两种方式进行通信：页面描述语言（Page Description Language，PDL）和打印机命令集。具体选择哪一种通信方式，必须根据不同的打印材料来决定。如果打印输出是全文字页面、图形和图像，那么就选择 PDL 通信方式；而如果打印输出是计算机制图，那么就选择图形语言接口方式。不过，很多制图程序中也可以提供 PDL 输出功能。因此，如果需要一个 PDL 接口，我们在通信方式上就有很多选择，但同时必须对打印输出材料进行仔细分析。

当计算机中的处理过程结束之后，打印机驱动负责将空心字、图形目标和图形转换成一个数据流，并输入到打印机中；打印机驱动具体实现了以下功能：在其他图像处理操作中进行数字半色调、重新分级、色彩数据变换和色彩外观调整等，这些功能都是为了实现最好的印刷质量。计算机中的数据压缩和打印机中的数据解压缩可以用来防止打印速度对数据通信速度的影响。进行内部数据处理的打印机中包含一个硬件格式程序板块，该格式程序板块中通常定义了打印机的所有规范；尤其是包含 PDL 接口的打印机更是如此。这种格式程序板块为打印机带来了很多优点，如加快了打印机的通信速度，缓解了主机的处理负荷，这一点对于复杂的文档处理非常重要。

随着打印文档复杂性的不断提高，实际的打印系统也越来越向满足用户需求的方向发展：即实现打印过程的可视性和可控制性、字体管理、快速回到软件应用程序和打印机配置。打印过程的可视性和可控制性取决于应用环境和操作环境的选择。各种字体（轮廓或位图）可能位于磁盘上，也可能位于计算机或打印机的只读存储器（ROM）中。为了加快打印速度，正在使用的空心字会被进行光栅化并保存在格式板块的随机访问存储器（RAM）或计算机的 RAM 中。不过，如果在打印过程中遇到了空心字，光栅化也会发生，那么打印速度就会变得非常慢，甚至慢到让人无法接受的程度。如果返回到应用程序的速度非常关键，那么在打印机中就必须包含各种格式的程序板块。因此，要想满足用户需求，就必须从系统的角度出发，对整个打印系统的配置进行分析（如计算机硬件、操作系统、应用程序、互连接口、打印格式模块以及打印机 CPU、存储器和字体保存信息）。

在打印彩色图像和复杂彩色阴影图形时，就会涉及到色彩搭配、颜色外观和色彩打印质量等问题。彩色打印机的配置包括：半色调规则、色彩搭配方案以及各种巧妙处理；这些巧妙处理是指根据目标是文字字符、图像或图形来进行标准化的色彩处理。如果输入设备和软件应用程序可以提供这些处理服务的话，那么打印过程就会变得更加复杂，而且每个彩色打印目标在打印之前都需要进行大量

的色彩处理。这种现象会严重影响打印质量，因此，检查整个图像处理链并消除大量色彩处理就变得非常重要了。彩色打印机配置选择的关键之处就是在打印质量和打印速度之间选择一个合适的平衡点。其中，半色调规则（可以使可见打印纹理最小化）和高印刷质量模式（要求反复印刷）会占有很多处理时间。对于彩色图像和图像来说，CRT 图像和硬拷贝之间的关系是彩色打印机配置选择的关键。在彩色图形中，通常采取牺牲色调精度的方法来实现色彩搭配或打印饱和度。在自然图像中，尤其是肤色图像中，色调精度更加重要，因此，必须在色调精度和色彩搭配之间选择一个更加合适的平衡点。有些软件和硬件提供商可以提供一些默认配置，这些配置可以根据打印内容来选择最佳的处理方式。如果需要非常精确的色彩控制，就要求色彩输入输出设备可以提供色彩重现机制，该色彩输入输出设备连接到一个 PC。这就涉及到了颜色管理的范畴了。

颜色管理

颜色管理主要是指如何使 3 种主要色彩设备（输入、显示、输出）在系统配置条件下实现相互协作；其中，技术要点就是数据表示方式。每个设备都包含一个内部色彩信息表示方式，该表示方式与表示或记录对象信息的本质直接相关。对于打印机来说，色彩信息表示方式就是指青色、洋红色、黄色和黑色（CMY，K）油墨；对于显示设备来说，色彩信息表示方式就是指红、绿、蓝（RGB）像素；而对于输入设备来说，色彩信息表示方式就是指 RGB 的数值。这些色彩内部表示空间称为“设备空间”，其体积是在三维彩色空间中定义的，该三维彩色空间可以被这些设备访问。为了在设备之间实现相互协作，这些内部空间会根据分析模型或三维查询表（LUT）来转换成一个设备独立空间。目前，实际中采用的是“CIE（Commision Internationale d'Eclairage）”比色空间，该空间是基于“CIE 1931 标准观察者”的；该空间使设备空间可以与人体色彩感觉标准对应起来。这些空间转换规则称为“设备规则”，而且常见的设备独立彩色空间称为“规则映射空间（Profile Connection Space，PCS）”。在这些空间转换过程中，我们可以发现，每个设备都可以映射到人体色彩空间的不同部分。例如，一个 CRT 显示的黄色就无法达到大多数彩色打印机上可以达到的饱和度。如果想得到满意的色彩饱和度，就必须进行大量的色彩处理，同时还要提供良好的视觉条件和用户的适应状态。解决这个问题的方法称为“颜

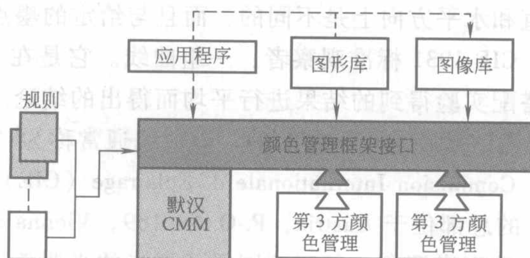


图 23-13　采用 ICC 规则的 ICC 颜色管理

注：基本单元包括：CM 框架接口、CMM、第三方 CMM 和规则（可能嵌入在文档中）

色管理方法（Color Management Methods，CMM）”（见图 23-13）。颜色管理系统的目标就是协调并完成这些操作。

因此，颜色管理系统的目的就是以最小的系统开销来提供最佳的色彩参考搭配、色彩加工以及色彩文件传输能力。颜色管理方案通常包含 3 种类型：点方案、应用方案和操作系统方案。点方案负责执行设备驱动中的所有处理操作，并严格符合系统要求。如果 CRT 需要进行颜色搭配，就必须提供 CRT 制造信息或者可视化校准工具来对 CRT 进行校准以满足驱动要求。应用方案中包含了各种设备规则库以及相关的 CMM，这种方案对于外围设备和应用软件提供商来说是透明的。操作系统方案嵌入在操作系统的功能模块中，这些系统可以提供一个默认的颜色搭配方法，但同时也允许使用提供商指定的 CMM。

尽管设备规则的生成过程涉及到很多直接的衡量标准，但要想得到满意的颜色搭配，就需要完成很多工作。在使用相同的视觉条件进行估计时，CIE 比色法特性决定了两种颜色的搭配程度。视觉条件完全相同是很少出现的，因此必须进行大量的适应性调整，称为“色彩外观变换”。一个简单的例子就是，彩色扫描仪中发光体的色度比 CRT 中白点的色度复杂得多；而 CRT 中白点的色度又比周围的视觉发光体白点的色度要复杂。另外，如前所述，不同的设备映射不同的颜色空间区域；也就是说，不同的设备具有不同的色彩范围。目标设备（如打印机）色彩范围之外的颜色必须变换到打印机色彩范围之内。如果源点设备与目标设备之间的动态范围不匹配，也可以采用这种色彩变换方法。能完成以上所有处理的技术是非常复杂的，而且这些技术都在提供商的专用 CMM 中形成了专利。

名词解释

精度：纸面上墨点之间的间距。通过单位长度上的墨点数来表示。该属性在垂直和水平方向上是不同的，而且与给定的墨点直径无关。

CIE 1931 标准观察者：一组曲线。它是在 1931 年通过对色盲观察者进行颜色搭配实验得到的结果进行平均而得出的结论。相关的光谱色彩光源是通过混合 3 种光谱激励来进行搭配的。该曲线通常称为“色彩搭配曲线”。

Commision Internationale d’Eclairage（CIE）：照明和色彩测试国际标准组织。CIE 的总部位于 Austria，P. O. Box169，Vienna a-1033。

数字半调色：基于相同尺寸墨点的半调色技术。用来模拟白纸和全彩色覆盖之间的灰色阴影。

灰阶控制：标识技术的内在调节特性。标识技术可以打印具有不同大小和密度的墨点。

半调色：通过改变彩色覆盖区域的大小来模拟连续色调的技术。通常是通过改变打印墨点的大小和相关密度来实现的。

H 和 D 曲线：感光材料的特征响应曲线。该曲线与曝光的光学密度相关。

分辨率：着墨点之间的间隔。通过调节该间隔可以实现墨点全覆盖；该间隔可以根据墨点大小来计算，它代表了打印机的基本性能。

饱和度：在颜色中用来描述无色轴上的彩色度。一种颜色达到饱和时是指该颜色没有消色差（无色）成分。

参考文献

[1] Cornsweet, T. N. 1970. *Visual Perception.* Academic Press, New York.

[2] Diamond, A. S. , ed. 1991. *Handbook of Imaging Materials.* Marcel Dekker, New York.

[3] Durbeck, R. C. and Sherr, S. 1988. *Hardcopy Output Devices.* Academic Press, San Diego, CA.

[4] Hunt, R. W. G. 1992. *Measuring Color*, 2nd ed. Ellis Horwood, England.

[5] Hunt , R. W. G. 1995. *The Reproduction of Color*, 5th ed. Fountain Press. England.

[6] Scharfe, M. 1984. *Electrophotography Principles and Optimization.* Research Studies Press Ltd. , Letchworth, Hertfordshire, England.

[7] Schein, L. B. 1992. *Electrophotography and Development Physics*, 2nd ed. Springer-Verlag, Berlin.

[8] Schreiber, W. F. 1991. *Fundamentals of Electronic Imaging Systems*, 2nd ed. Springer-Verlag, Berlin.

[9] Ulichney, R. 1987. *Digital Halftoning.* MIT Press, Cambridge, MA.

[10] Williams, E. M. 1984. *The Physics and Technology of Xerographic Processes.* Wiley-Interscience, New York.

备注

读者还可以参考以下资料或相关组织：

[1] *Color Business Report* ：由 Blackstone Research Associates 出版，该机构位于 MA 015690345, Uxbridge, P. O. Box 345。内容涉及到与色彩、计算机和复印等要点相关的各种行业。

[2] 国际色彩联盟（International Color Consortium, ICC）：该联盟的基本成员包括：Adobe Systems Inc.、Agfa-Gevaert N. V.、Apple Computer, Inc.、Eastman Kodak Company、FOGRA（Honorary）和 Microsoft Corporation；*Hewlett-Packard Journal* 1985. 36（5）；1988. 39（4）（该杂志主要介绍热喷墨技术）。

[3] *Journal of Electronic Imaging*：由 IS&T 和 SPIE 联合出版。主要介绍图像数据的获取、显示、通信和存储，以及硬拷贝输出、图像可视化和相关图形主题。该资料是目前打印机中色彩处理和数字半调色研究的主要参考来源。

[4] *Journal of Imaging Science and Technology*：IS&T 的官方出版物。其主题涵盖了各种成像技术，包括卤化银技术和计算机打印技术。

[5] 国际光学工程协会（SPIE）：位于 Washington 982270010, Bellingham, P. O. Box 10。该协会联合 IS&T 一起主办了各种关于电子成像的会议，并出版了相关的会议学报。

[6] *Hardcopy Observer*：由位于 MA 02160, Newtonville, P. O. Box 9143 的 Lyra Research Service 机构出版的一本行业观察杂志月刊；该月刊提供了家用和办公打印机行业的发展趋势。

[7] 成像科学和技术协会（Society for Imaging Science and Technology, IS&T）：位于 VA 22151, Spring-

field，7003 Kilworth Lane。电话（703）642 9090，传真（703）642 9094。该协会主办了各种关于成像和打印技术的技术会议，并出版了会议学报、书籍和“*Journal of Electronic Imaging*”、“*Journal of Imaging Science and Technology*”和“*IS&T Reporter*”等期刊。

[8] 信息显示协会：位于 CA 92705-5421，Santa Ana，Ste 82，1526 Brookhollow Drive。电话（714）545 1526，传真（714）545 1547。该协会每年与 IS&T 共同主办关于彩色成像技术的年会。

第 24 章　数据存储系统

Jerry C. Whitaker

24.1　引言

如今，各种系统之间传输的数据量越来越大，因此就带来了数据管理方面的需求；这样，越来越多的服务器——无论是 PC、UNIX 工作站、微型计算机，还是超级计算机——都可以担任信息提供者和管理者的角色。各种互联系统的数量正在飞速增长，这主要得益于客户-服务器计算机模型的广泛应用。硬盘存储在推动互联系统的发展中起到了关键作用，因为不断增长的海量数据需要大量的存储空间。同时，随着数据系统的发展，我们要求存储系统提供商的存储产品不仅可以存储海量数据，而且还可以对海量数据进行快速访问并支持多用户同时访问。这样的存储系统还必须是安全的，而且可以确保数据永远不会丢失，而且网络系统用户可以随时访问。

24.2　独立磁盘冗余阵列系统

常见的海量数据快速可靠访问的实现方法就是将多个磁盘驱动器组合在一起，形成一个磁盘阵列，称为“独立磁盘冗余阵列（Redundant Arrays of Independent Disk，RAID）”子系统。最简单的 RAID 系统是由一组 5～6 个磁盘驱动器构成的，这些磁盘驱动器装在一个盒子中，并连接到一个控制器板块上。RAID 控制器可以控制这些磁盘驱动器协调一致地进行读写操作，如同控制器在控制单个磁盘驱动器一般；我们还可以将磁盘阵列看做成一个驱动器或虚拟驱动器。驻留在主系统中的 RAID 管理系统负责对存储在 RAID 子系统中的数据进行管理。图 24-1 给出了一个典型的 RAID 配置示例。

1. RAID 组成

尽管 RAID 系统中包含多个驱动器配置，但是 RAID 子系统中的各个磁盘驱动器对于用户来说都是透明的；即使 RAID 子系统可以无限大，但 RAID 子系统本身就是一个虚拟驱动器。这种虚拟驱动器是通过 RAID 管理软件在主操作系统中形成的。RAID 管理软件不仅可以使系统将 RAID 单元当作单个驱动器一样进行访问，而且可以使子系统的配置最大程度上满足主系统的需要。

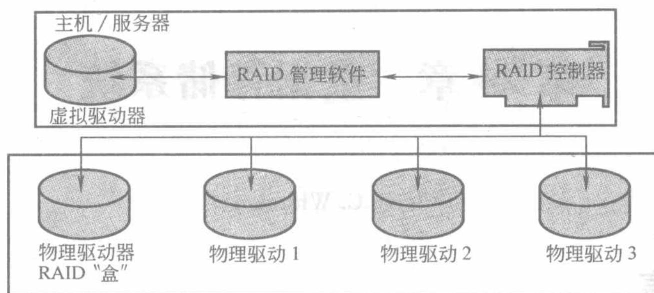


图 24-1 一个典型的 RAID 配置示例（来源：Adapted from Heyn, T. 1995. The RAID Advantage. Seagate Technology Paper, Seagate, Scotts Valley, CA.）

RAID 子系统可以在性能、最大容量、容错能力等方面进行优化。根据这些可优化参数，可以对不同的 RAID 等级进行定义和标准化。RAID 包含 6 个标准等级，根据性能、冗余度和主系统要求的其他属性，这些等级分别记为 RAID 0、1、2、3、4、5 级。

RAID 控制板是一个硬件单元，它是磁盘阵列的主架构。RAID 控制板不仅可以向磁盘阵列中的指定驱动器转发输入输出（I/O）命令，还可以为每个独立的驱动器提供物理连接，以便对其进行读写。该控制板还可以监控阵列中每个驱动器的完整性，以便预测错误或故障磁盘驱动器的数据传输（这个特征称为“容错能力”）。

2. RAID 等级

RAID 0 ~ 5 标准为用户和系统管理者提供了很多配置选择，这些选择允许磁盘阵列选择合适的应用环境。这些配置中每一种配置的目的都集中在如何使以下几个方面达到最佳：

1）容量；

2）数据有效性；

3）性能；

4）容错能力。

（1）RAID 0 级配置

配置级别为 RAID 0 级的磁盘阵列是一个性能经过优化的磁盘阵列，但同时也牺牲了容错能力和数据完整性作为代价。RAID 0 级配置是通过一种称为“数据分解”的方法来实现的。RAID 0 级阵列磁盘驱动器（虚拟驱动器）中的数据是以数据条的形式存储的。一个典型的阵列中可以包含任意多个数据条，具体数量通常为阵列中驱动器数目的整数倍。例如，一个 4 驱动器阵列的配置中可以包含 12 个数据条（每个指定的驱动器中包含 3 个数据条）。其中，数据条 0、1、2

和 3 分别位于驱动器 0、1、2 和 3 中；数据条 4 也位于驱动器 0 中，但它与数据条 0 的位置不同；同样，数据条 5～7 分别位于驱动器 1、2 和 3 中的相应位置上。剩下的 4 个数据条的位置依此类推，这样数据的存储方式就是相同的，如图 24-2 所示。事实上，在给定的 RAID 子系统中，驱动器可以形成任意多个数据条，如在两个驱动器上可以形成 200 个数据条，其性能与 50 个驱动器中的 50 个数据条相同。不过，大多数 RAID 子系统中一般只包含 3～10 个数据条。

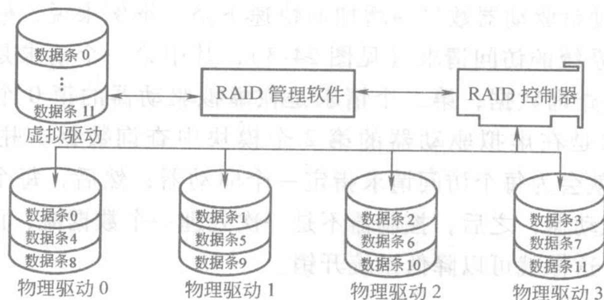


图 24-2　在 RAID 0 级配置中，虚拟驱动器包含多个数据条

注：各个连续的数据条均匀分布在相邻的物理驱动器中，形成了一个数据条链。

（来源：Adapted from Heyn，T. 1995. The RAID Advantage. Seagate Technology Paper，Seagate，Scotts Valley，CA.）

RAID 0 级配置是一个性能经过优化的配置，这是因为数据分解使得阵列可以同时对多个驱动器进行访问。换句话说，由于数据可以扩展到阵列的多个驱动器中，这样数据访问就不会被限制在单个驱动器中，因此访问速度就提高了。这对于对超大型文件的访问非常重要，因为数据访问可以跨越多个驱动器，如同在一个驱动器中对各个数据条进行访问。

RAID 0 级配置中最底层数据条的容错能力会有所下降，从而增加了数据丢失的风险，这是因为没有有效空间来存储冗余数据了。如果 RAID 0 级配置中的某个驱动器出现故障，就无法恢复丢失的数据了，而在其他 RAID 配置则可以恢复。

（2）RAID 1 级配置

RAID 1 级配置采用了一种称为“磁盘镜像”的技术，该技术用来确保数据的可靠性或高容错能力。RAID 1 级配置同样可以提高访问性能，但提高性能和容错能力的同时，也会牺牲驱动器的有效容量作为代价。在 RAID 1 级配置中，RAID 管理软件指示子系统控制器在阵列中各个驱动器（镜像驱动器）中存储数据。换句话说，相同的数据会被复制并保存在不同的磁盘上，以确保阵列中某个驱动器出现故障时，其他地方的数据仍然有效。事实上，所有的驱动器都有可能出现故障，但只有镜像驱动器不会，而且保存在 RAID 子系统中的数据也会保持完整无缺。RAID 1 级配置中可以包含多个镜像驱动器，每个镜像驱动器的容量

各不相同。通常，构成一个镜像驱动器的各个驱动器容量是相同的。如果一个镜像驱动器中的各驱动器容量不相同，那么 RAID 1 级配置中镜像驱动器的容量就会受到驱动器中最小容量驱动器的限制，从而也会造成多个驱动器的有效容量受到影响。

如果冗余数据均匀分布在 RAID 子系统所有的镜像驱动器上，这样就可以提升系统的访问性能。事实上，访问请求的数量和等待状态的总次数都会随着 RAID 子系统中硬盘驱动器数量的增加而快速下降。举例来说，假设现在有 3 个对 RAID 1 级子系统的访问请求（见图 24-3），其中第一个请求是在虚拟驱动器的第 1 个模块中查询数据；第二个请求是在虚拟驱动器的第 0 个模块中查询数据；第三个请求是在虚拟驱动器的第 2 个模块中查询数据。驻留在主机中的 RAID 管理软件就会为每个访问请求指定一个驱动器；然后，每个请求指令就会发送到相应的驱动器；之后，控制器不是一次处理一个数据流，而是几乎同时发送 3 个数据流，这样就可以降低系统开销。

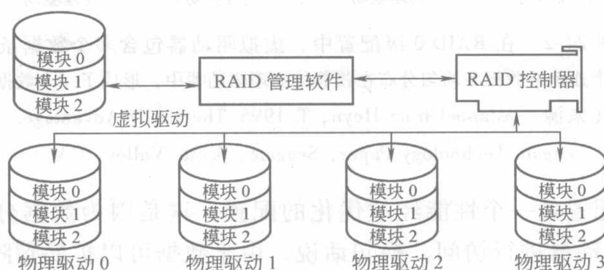


图 24-3 RAID 1 级配置可以通过物理硬盘之间的数据复制（镜像）来提供很高的数据可靠性；另外，由于 RAID 管理软件允许在多个驱动器之间同时分配多个访问请求，因此 I/O 性能也会得到提升（来源：Adapted from Heyn，T. 1995. The RAID Advantage. Seagate Technology Paper，Seagate，Scotts Valley，CA.）

（3）RAID 2 级配置

RAID 2 级配置很少出现在商业应用中，但也是一种确保子系统驱动器中数据不出现问题或故障的一种方法。RAID 2 级配置中的容错能力是由汉明纠错码（Error Correction Code，ECC）生成的，汉明 ECC 通常用在调制解调器和固态存储器设备中，用来保证数据的完整性。ECC 利用一定的算法将存储在虚拟驱动器指定模块中的各种数据数值制成表格，该算法可以生成校验和。如果需要，该校验和就会附加在数据模块的末端，用来校验数据的完整性。

当数据从驱动中读取出来时，ECC 表格就会重新计算，而且指定数据模块的校验和就会被读取并与最新的表格进行比较。如果结果匹配，就说明数据是完整的；如果存在差异，就必须利用第一个校验和或更早的校验和作为参考点来重新计算丢失的数据，如图 24-4 所示。

ECC 的这种原理与驱动器自身使用的 ECC 技术是不同的。与其他 RAID 实现方式的性能相比，RAID 2 级配置阵列中数据存储的拓扑格式性能是非常有限的，这就是为什么 RAID 2 级配置在商业应用中很少见的原因。表 24-1 所示为纠错码（ECC）示例。

表 24-1　纠错码示例

假设数据分量存储在“HELLOTHERE”中，数据中的每 10 bit 进行校验和计算										
存储的数据	H	E	L	L	O	T	H	E	R	E
数字表达式	71	69	76	76	79	84	72	69	82	69
校验和格式	$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	$x7$	$x8$	$x9$	$x10$
乘数值	72	138	228	304	395	504	504	414	738	690
[校验]求和	72	+138	+228	+304	+395	+504	+504	+414	+738	+690
										=3987

因此，存储在驱动器上的数据为：72 69 76 76 79 84 72 69 82 69 3987

当数据从驱动器中读取出来时，对数据分段将进行相同的计算，最新计算的校验和会与之前存储的校验和进行比较，以验证数据的完整性

来源：Adapted from Heyn, T. 1995. The RAID Advantage. Seagate Technology Paper, Seagate, Scotts Valley, CA.

（4）RAID 3 级配置

RAID 3 级配置实际上是 RAID 0 级配置的改进型，RAID 0 级配置牺牲了一定的容量；而对于相同数量的驱动器，RAID 3 级配置可以实现更高的数据完整性和容错能力。RAID 3 级配置也采用了 RAID 0 级配置中的数据分解方法，不过阵列中的所有驱动器只有一个驱动器没有进行数据分解；该驱动器存储了子系统中所有驱动器用来维持数据完整性的奇偶信息。奇偶驱动器自身也被划分成多个奇偶驱动条，每个奇偶驱动条用来存储阵列中对应数据条的奇偶信息。这种方法可以通过对所有驱动器进行并行或同时读写来实现很高的数据转移性能，而且一旦驱动器出现故障，该方法还保留了数据重构方法，以维持系统数据的完整性。图 24-4 详细说明了这种方法的原理。对于大型连续文件的实时访问来说，RAID 3 级配置是一种很理想的配置。

存储在专用驱动器中的奇偶信息是利用专用 OR 函数来计算的。利用专用 OR 函数并结合 RAID 中的一系列数据条，任何丢失的数据都可以很轻松地恢复。如果阵列中的某个驱动器出现故障，丢失的信息也可以通过类似于求解方程式中单个变量的方式来恢复。

（5）RAID 4 级配置

RAID 4 级配置的概念与 RAID 3 级配置非常相似，但 RAID 4 级配置更加注重不同应用的性能，如数据库文件和大型连续文件；另外，RAID 4 级配置具有更大的数据条长度，通常为两个模块，该模块允许 RAID 管理软件在控制磁盘时

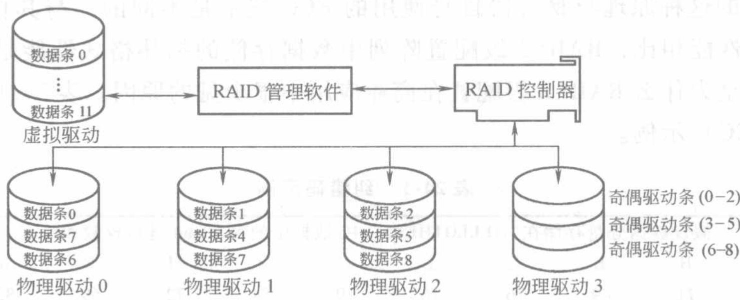


图 24-4 在硬盘驱动中数据条的使用方式上，RAID 3 级配置与 RAID 0 级配置非常相似

注：除了这些数据条之外，系统指定了一个专用驱动器来存储奇偶信息，以维持 RAID 子系统中数据的完整性

（来源：Adapted from Heyn，T. 1995. The RAID Advantage. Seagate Technology Paper，Seagate，Scotts Valley，CA. ）

比 RAID 3 级配置更加独立（RAID 3 级配置是统一控制磁盘）。这样，RAID 4 级配置在拥有与 RAID 3 级配置相同高数据吞吐量的同时，还加快了密集读取应用中数据访问的速度（见图 24-5）。

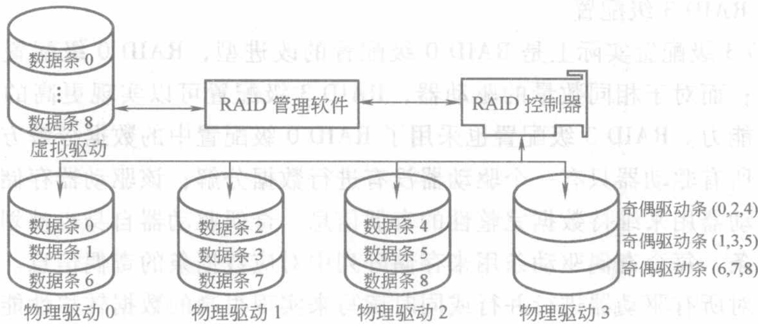


图 24-5 RAID 4 级配置的基础是 RAID 3 级配置

注：为存储数据条配置了不连续的奇偶信息条，这样，就可以实现独立的磁盘管理，非常适合密集读取环境

（来源：Adapted from Heyn，T. 1995. The RAID Advantage. Seagate Technology Paper，Seagate，Scotts Valley，CA. ）

RAID 4 级配置的不足之处在于奇偶驱动器上存在瓶颈。当数据写入阵列时，写操作的奇偶编码方案就会变得比其他 RAID 拓扑配置长很多。这一点或多或少会影响 RAID 4 级配置在密集读取应用中的使用。因此，如同 RAID 3 级配置，RAID 4 级配置在商业应用中也很少出现。

（6）RAID 5 级配置

RAID 5 级配置是最后一种常见的 RAID 配置，而且也可能是应用最多的配

置。RAID 5 级配置通过把奇偶信息条分布在一系列硬盘驱动器上的方式，将 RAID 4 级配置的写入瓶颈降到了最低程度。这样，RAID 5 级配置就可以缓解单个驱动器上的集中写入操作，从而可以提高整体系统的性能（见图 24-6）。

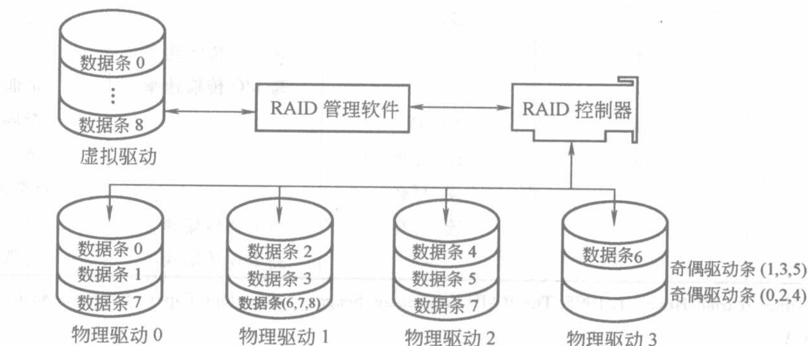


图 24-6 RAID 5 级配置通过分布式的奇偶信息条策略克服了 RAID 4 级配置的写入瓶颈，这样可以更好地分配 RAID 驱动器上的写入操作，从而提高系统性能。（来源：Adapted from Heyn，T. 1995. The RAID Advantage. Seagate Technology Paper，Seagate，Scotts Valley，CA.）

RAID 5 级配置降低奇偶写入操作瓶颈的方式是相对简单的。这种瓶颈风险不是由阵列中的某个驱动器单独承担，而是阵列中的所有驱动器共同承担。这种分布式风险承担机制释放了集中在单个驱动器上的写入操作，从而可以提高子系统的整体吞吐量。RAID 5 级奇偶编码方案与 RAID 3 级和 4 级配置相同；当某个驱动器出现故障时，可以确保系统具有恢复丢失数据的能力。这种情况只有当某个驱动器上没有奇偶信息条可以存储相同驱动器上的数据条信息时才会出现。换句话说，任何数据条的奇偶信息必须始终位于另一个驱动器上，而不是数据所在那个驱动器。

（7）其他 RAID 配置

各个 RAID 配置提供商也开发了其他不常见的 RAID 配置，包括：

1）RAID 6 级配置：强调超高数据完整性；

2）RAID 10 级配置：强调高 I/O 性能和极高数据完整性；

3）RAID 53 级配置：合并了 RAID 0 级和 3 级，以实现统一的读写性能。

各 RAID 级别的属性如表 24-2 所示。

（8）定制 RAID 系统

RAID 技术的最大优势可能就是包含了各种适合用户和系统设计者的配置方案；RAID 可以根据环境要求和应用需求来定制专用的阵列子系统。各种配置选择方案可以满足各种特定的应用要求，如表 24-2 所示。但是，定制配置不会形成一个专门的配置级别，其中必须考虑驱动器模型、容量和性能级别和有效连接选择等因素。

表 24-2 RAID 属性概括

RAID 级别	容　量	数据有效性	数据吞吐量	数据完整性
0	高	读/写高	高 I/O 传输速率	
1		读/写高		镜像
2	高		高 I/O 传输速率	ECC
3	高		高 I/O 传输速率	奇偶
4	高	读高		奇偶
5	高	读/写高		奇偶
6		读/写高		双奇偶
10		读/写高	高 I/O 传输速率	镜像
53			高 I/O 传输速率	奇偶

来源：Adapted from Heyn, T. 1995. The RAID Advantage. Seagate Technology Paper, Seagate, Scotts Valley, CA.

名词解释

光纤判决信道环（Fiber Channel-Arbitrated Loop，FC-AL）：一种高速接口协议，可以提供高速数据传输和大量连接设备，而且可以远程操作应用光纤和铜电缆连接的设备。

独立磁盘冗余阵列（RAID）：硬盘驱动器的一种配置，可以为海量存储提供管理软件，具有高容量、高速率和高可靠性。

RAID 等级：RAID 单元的标准配置，其目的是实现特定的目标，如最高的可靠性或最大的速率。

单连接器附件（Single Connector Attachment，SCA）：某种接口标准的电缆规范，该标准通过合并数据和电源信号来形成一个共同的标准化端口，简化了设备间的互连互通。

小型计算机系统接口（Small Computer Systems Interface，SCSI）：用来将多个设备连接到计算机系统的电缆和软件协议。这些设备可能是计算机内部的，也可能是外部的。SCSI 包含很多类型。

数据分解：硬盘存储组织技术，该技术将数据存储在多个物理设备上，从而提高读写速度。

热校准：硬盘驱动器的内部整理功能，用来保持磁盘表面磁点的合适排列。

虚拟驱动器：计算设备的一种工作状态，其中存储器的组织方式可以实现特定的目标，虚拟驱动器通常不是真正的物理实体。例如，计算机中的 RAM 可以作为一个物理驱动器出现，或者两个或多个硬盘也可以作为一个物理驱动器出现。

参考文献

[1] Anderson, D. 1995. Fiber channel-arbitrated loop: The preferred path to higher I/O performance, flexibility in design. Seagate Technology Paper No MN-24, Seagate, Scotts Valley, CA.

[2] Heyn, T. 1995. The RAID Advantage. Seagate Technology Paper, Seagate, Scotts Valley, CA.

[3] Tyson, H. 1995. Barracuda and elite: Disk drive storage for professional audio/video. Seagate Technology Paper No SV-25, Seagate, Scotts Valley, CA.

备注

关于硬盘驱动器设计和制造的技术不在本章介绍的范围之内，而且由于大多数应用中都将磁盘驱动器当作一个可操作器件来使用（如果你愿意，可以当作是一个黑匣子），因此关于磁盘驱动器的最佳最新信息来源就是驱动器制造商，而技术应用注意事项和详细的产品规范就不大需要了。

第25章 光学存储系统

Praveen Asthana

25.1 引言

可记录光学磁盘驱动器技术可以为不断增长的可移动存储需求提供有效解决方案。相对于附加的存储介质（价格相对很便宜，如光盘）来说，光学磁盘驱动器可提供的存储容量是非常有限的。这种成效比很高的存储性能在存储集中的现代计算机应用系统中非常受欢迎，如台式印刷系统、计算机辅助设计/计算机辅助制造（CAD/CAM）或多媒体制作。

25.2 光度头

光度头的作用就是将激光束投射到光盘上，并将激光束聚焦在一个衍射点上，然后将从光盘上读出的信号信息传输到数据和伺服检波器中。

无论记录技术是磁性材料、烧蚀 WORM 还是相变材料，激光二极管都是光学存储器中的关键器件。早期的光学驱动器采用红外激光，其发射波长为 780nm 或 830nm。最新的光学驱动器采用红色激光，发射波长为 690nm。激光被公认在 40mW 功率范围内具有最大的连续输出功率，而且可以根据折射率进行引导，以确保良好的波阵面质量。

在激光二极管中，光是从一个“切面”上发射出来的，该切面是折射率引导激光光导区域的断开面。切面必须足够小（通常为几微米），以便光从切面上发射出来时产生衍射，输出的光束具有大量的发散角。在很多商用激光二极管中，切面的宽度（例如，类似于 PN 结面的宽度）远超过其高度（或者在与 PN 结面垂直的方向），这样在激光二极管 PN 结面的水平和垂直方向上的发散角就不相同了，而且远离激光二极管一段距离时激光束的空间规律就可以忽略不计了。

图 25-1 给出了磁-光驱动器光度头的基本构造示意图。在该结构中，激光二极管安装在一个散热器中，以便及时驱散热量。激光二极管输出的光是通过透镜 1 进行校准的。其中，一个类似棱镜的光学器件称为“折射器”，用来降低激光束的椭圆率。然后，激光束就会穿过一个偏振分光器，该分光器会将部分（30%）光束反射到检波器中，之后这部分光束就会投射到磁盘上。

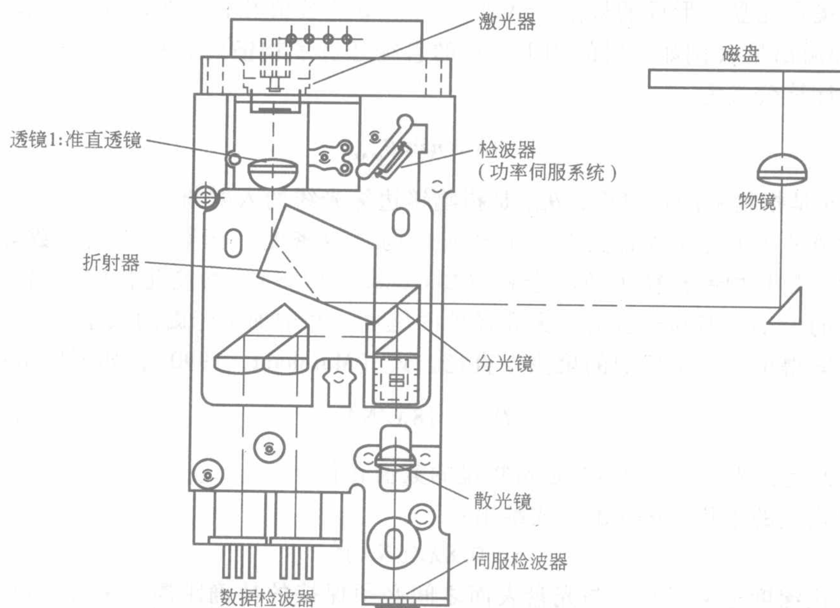


图 25-1　磁-光驱动器中光度头的基本构造示意图

注：该构造图是一个分光设计，由固定器件和可移动系统构成，可移动系统包含分光器和物镜，物镜用来将光聚焦到光盘上

（来源：Asthana，P. 1994. Laser Focus World，Jan，P. 15. Penwell Publishing Co，Nashua N. H. Used by Permission.）

激光二极管输出的光会在平行于 PN 结的方向上产生偏振（称为“P 偏振”）。发射光的 P 偏振分量密度与 S 偏振分量密度的比值大于 25:1。这种偏振分光器可以输出 70% 的 P 偏振光和 100% 的 S 偏振光。被反射的光会入射到光检波器中，该检波器是功率伺服系统中的一部分，伺服系统用来使激光二极管的功率保持在一个恒定的状态。如果没有功率伺服系统，激光二极管的功率会随着其中 PN 结的变热而产生波动，这会对系统的读取性能产生不利影响。

分光器发射出来的光束会通过一个旋转镜（90°），称为“偏光器”，偏光器安装在一个可移动的传动装置上。在光盘磁轨寻找操作过程（寻轨）中，该传动装置可以在光盘上呈放射状地移动。被旋转镜反射出来的光会入射到一个物镜中（同样安装在传动装置中），该物镜可以将光束聚焦到光盘上。这种激光光度头设计称为“分光设计”；其中，激光二极管和大多数光学器件都是固定的，而物镜和偏光器是可以移动的。在早期的光学驱动器设计中，整个光度头都安装在一个传动装置中，而且在寻轨过程中是一起移动的；这种设计使寻轨时间变得很长（约 200ms），因为传动装置的负载过大。在分光设计中，由于相干光的校准率很高，减小了传动装置的负载，因此其寻轨速度就快得多。

物镜在光盘上形成的聚焦点大小是由物镜的数值孔径（Numerical Aperture，NA）和溢出量（例如入射的相干光束的直径超过了物镜的孔径）决定的。数值孔径的计算公式为

$$NA = n\sin\theta_{max}$$

式中 n 是指物镜的折射率；θ_{max}是指物镜边缘光线的入射角。

物镜的入射光束通常具有一个高斯电场（或密度）函数；焦点函数等于入射密度函数和物镜孔径函数的卷积（Goodman，1968）。物镜孔径的溢出会减小聚焦点的大小，但同时也会丢失孔径外的能量，并增加了旁瓣的大小。溢出量的优化方案得出了一个近似的聚焦点直径公式（Marchant，1990），如下所示：

$$D = 1.18\lambda/NA \tag{25-1}$$

式中，λ 是指光的波长；NA 是指物镜的数值孔径。

聚焦点的聚焦深度 z 由下式给出：

$$z = 0.8\lambda/(NA)^2$$

聚焦深度定义了物镜与光盘表面之间必须保持的精确距离。聚焦深度越小，系统对介质倾斜度的允许偏差就越小，聚焦伺服系统的工作就越难。因此，可以通过增加物镜的 NA 来减小聚焦点的大小（通常聚焦点越小，存储密度越大），但是当 NA 超过 0.6 时，这种方法就变得不可行了。对于光盘上反射回来的光，物镜也可以看做是一个聚光镜。这种反射光通常用于伺服系统和读取操作，其中包含了光盘上的读取信息。反射光会沿着入射路径到达固定的光学器件。分光器 1 在伺服检波器和数据检波器的方向上反射了所有的 S 偏振光和 30% 的 P 偏振光。

不过，被分光器发射的部分光束会被准直透镜聚焦回到激光二极管的断面上。即使在磁-光介质中这种反馈光的总量也不会超过总输出光束的 7%，但反馈光束会在激光二极管中导致很多问题。光反馈会造成激光二极管的发光模式出现随机跳跃，从而造成输出的光束振幅随机波动。这种振幅干扰是一个非常严重的问题，因此必须采取措施来控制这种振幅干扰（如注入高频电流）或 HFM（Arimoto，1986）。通常提高 HFM 电流会减小干扰，但在实际中，注入电流不能任意放大，因为它会破坏计算机附件中对辐射的限制。光反馈还会降低激光二极管的门限并提高功率-电流（PI）曲线的斜率，但这两个问题并不是主要的。

分光器 1 的反射光会进一步被分光器 2 分离成伺服分量和数据分量。分光器的反射光会入射到数据检波器中。对于磁-光回读来说，两个检波器采用的技术称为“差分检波”（例如，检测两个入射信号之间的差别）。穿过分光器 2 的光束会入射到特殊的多元伺服检波器，用来产生伺服信号。这种信号产生机制就是我们接下来要讨论的主题——伺服系统。

1. 伺服系统

伺服系统用来将激光点准确聚焦到光盘的任意磁轨上，并根据需要将激光点移动到光盘的任意磁轨上。光盘上的高磁轨密度（18000 磁轨/in）要求激光点的位置必须控制在 1μm 的若干分之一之内。为了在整个光盘表面区域内自由移动，传动装置必须很长；但是这样的传动装置太大的话，在光盘旋转时就不能快速响应磁轨位置的变化了。因此，一个合适的传动装置是由简单传动装置和精致传动装置构成的，用来控制激光束在光盘上的辐射位置。精致传动装置（质量很轻）可以在有限的范围内快速变化聚焦点的位置；而简单传动装置的响应速度较慢，但是它的移动范围更广，主要用于较长的寻轨操作。光盘中有一个连续的螺旋槽（如同唱片中的音轨），用来提供磁轨的相对位置信息。

除了激光寻轨和查找之外，无论光盘如何转动，光驱中的激光点都必须在光盘上保持最佳的聚焦效果（如果光盘有一些倾斜或歪曲，激光头就必须进行垂直移动）。为了实现这个目标，当光盘旋转时，物镜必须根据光盘表面的轴向运动来不断校正位置。物镜的位置是由伺服系统控制的。

图 25-2 给出了光驱伺服控制系统的模块结构图（包含了寻轨和聚焦过程）。返回的光束包含了聚焦点的各种信息，该信息之后由伺服检波器处理。从检波器得到的反馈信号确保了系统维持对光束的控制。

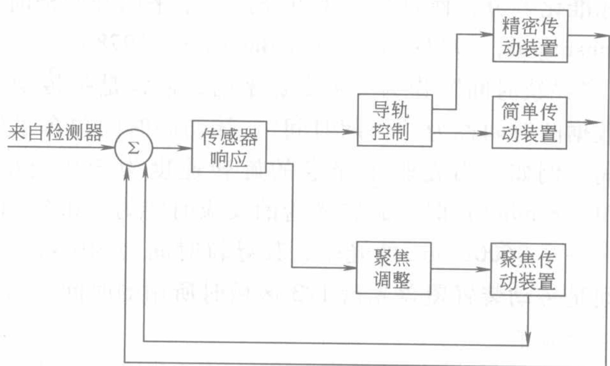


图 25-2 光驱伺服控制系统的模块结构图

聚焦控制系统需要一个反馈信号，该信号精确指明了聚焦误差的大小和方向（Braat and Bouwhuis，1978；Earman，1982）。为了生成一个聚焦误差信号，通常采用散光镜来将光束聚焦到一个四分检波器；其中，部分光束来自于光盘上的反射光。在聚焦很好的情况下，聚焦点是均匀分布在四分检波器的 4 个区域中的（见图 25-3）。不过，如果物镜不是处于聚焦状态，那么检波器上的聚焦点就会成椭圆形，这是由散光镜的光学特性造成的。检波器 4 个区域中光束分布的不均匀就会产生“聚焦误差信号（Focus Error Signal，FES）”；该信号会根据光线等级进行标准化，以确保光线与激光功率和光盘反射性无关。聚焦传动装置是由物镜构成的，该物镜的位置通过偏转线圈进行调整；其中，线圈绑在物镜四周以减小动量，同时恒久磁性也可以保持不变；而物镜可以通过一个滑动插销或弹性曲钉来固定。这种设计中的要点是活动范围、加速

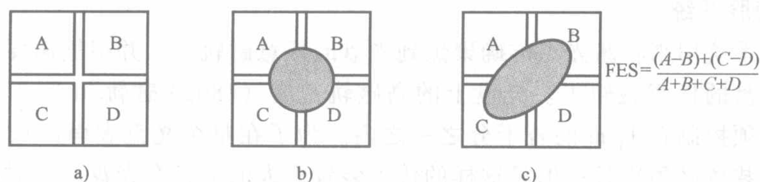


图 25-3 聚焦控制系统

a）四分检波器 b）通过散光镜聚焦在检波器四分区域的聚焦点（圆形表示物镜处于聚焦位置） c）物镜没有处于聚焦位置时在检波器四分区域上形成的聚焦点

度、无谐振和散热等特性。

一旦聚焦点聚焦在有效面上，聚焦点必须准确找到所需的磁轨并保持相应的位置，这就是寻轨伺服系统的功能。这时，通常可以采用产生聚焦信号的相同四分检波器来产生“寻轨误差信号（Tracking Error Signal，TES）”。从光盘反射回来的光束中包含了一阶衍射分量，这些衍射分量的密度取决于磁轨上聚焦点的位置，并沿着四分检波器的某个轴线逐渐下降。TES 是检波器两个半区上产生电流的标准化差分，而且当聚焦点经过一个平面和凹槽时，这种差分会达到一个峰值（Mansuripur，1987；Braat and Bouwhuis，1978）。

“寻轨时间”作为一种专业术语，通常是指传动装置的实际移动时间；而到达数据的时间称为“访问时间”，该访问时间包含了旋转光盘的反应时间和寻轨时间。例如，当光驱中光盘的旋转速度为 3600 转/分（Revolution Per Minute，RPM/(r/min)）时，旋转光盘的反应时间为（0.5×(60×3600)）s，即 8ms。因此，一个 3600r/min 的光驱，其寻轨时间为 30ms，访问时间为 38ms。标准寻轨时间是传动装置覆盖光盘 1/3 区域时所用的时间，这个标准是从早期的光驱中得到的经验值。

精确定位旋转光盘径向运动和轴向运动的功能与聚焦和寻轨传动装置的质量直接相关。为了防止因为振动、摇动和介质振摆而产生误差，伺服系统必须具有很高的带宽。限制伺服系统高带宽的因素通常是传动装置的共振模式。当传动装置的尺寸越小时，共振的频率就会越高，从而可达到的带宽就越高。光驱中的伺服系统还面临其他的挑战，因为它还必须处理可移动介质，这些介质在不同的光盘中其特性也各不相同（如介质倾斜）。

2. 光学记录和读出通道

图 25-4 给出了光驱的功能模块示意图。其中，SCSI 控制器负责处理光驱与主系统之间的信息流（包括命令）。光驱控制器是数据路径上的关键控制器，负责解释从 SCSI 控制器传过来的命令，并引导数据穿过随机存取存储器（RAM）到达写入通道，或者引导数据从读取通道输出。光驱控制微处理器单元通过逻辑门阵列来控制光驱中的所有功能单元，这些功能单元包括伺服控制、轴动引擎、

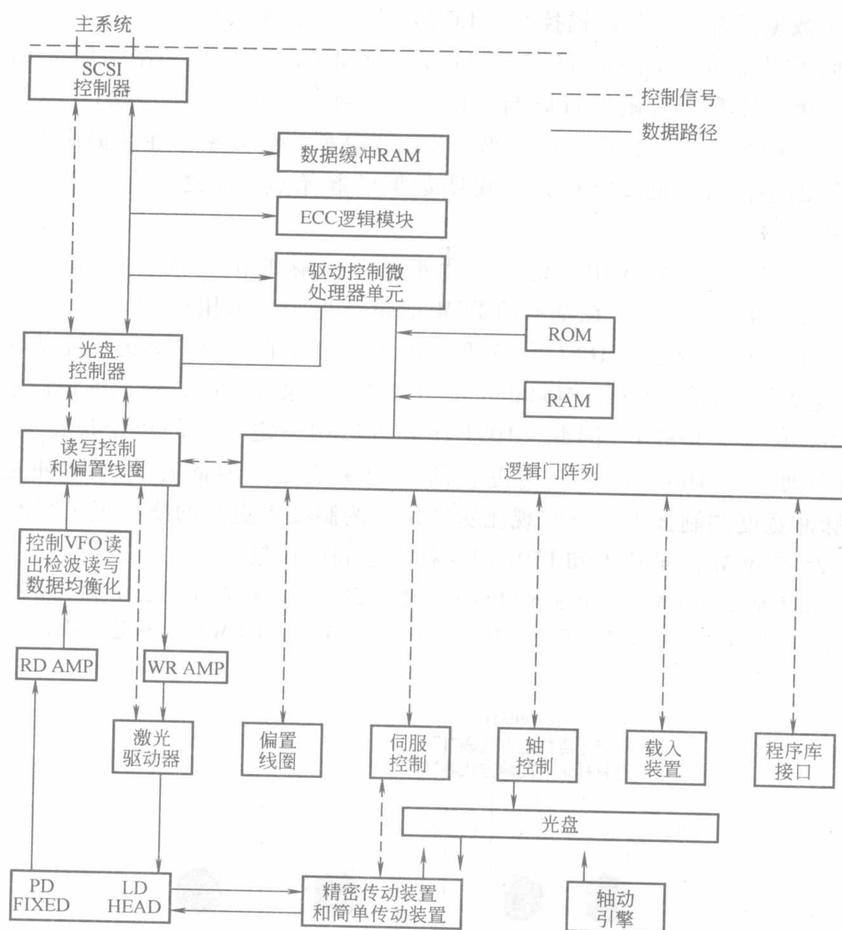


图 25-4 光驱的功能模块示意图

注：该图描述了关键的电气功能和接口。其中的缩写词分别为：VFO 是指可变域振荡器，用于数据同步；RD AMP 是指读出信号放大器；WR AMP 是指写入信号放大器；PD FIXED 是指光电检波器；LD HEAD 是指激光二极管

传动装置、激光驱动器等等。

通过 SCSI 输入到光驱中的数据首先分解成固定大小的分段（例如，512KB 或 1024KB 长），然后保存在数据缓冲 RAM 中。磁-光、相变和 WORM 驱动器可以看做是一种“固定模块结构”技术；在该技术中，数据的记录模式与硬盘非常类似（Marchant，1990）。光盘中的所有数据可以在任何位置以任何顺序排列。举例来说，目前的 CD 可记录驱动器就是一个无固定模块结构的例子（因为该结构的根源在 CD 音频中）。在常见的 CD-R 光驱中，输入数据是顺序记录的（类似于磁带），数据可以具有任意的连续长度。

每个数据段都会添加纠错控制（ECC）字节。光驱采用了 Reed-Solomon 码，该纠错码可以将错误率降低 1E-5 至 1E-13（Golomb，1986）。在添加了 ECC 信息之后，就可以利用有限运行周期（RLL）调制码（Tarzaiski，1983；Treves and Bloomberg，1986）来对数据进行编码，从而提高检测的效率。在数据流中，可以插入特殊的字符（如同步字符）来说明数据的起始点。总之，存储用户数据的开销约占 20%。

目前，在大多数光驱中，记录过程都是基于脉冲位置调制（Pulse Position Modulation，PPM）技术。在基本的 PPM 记录过程中，采用标记来表示位“1”，空白表示位“0”。因此，1001 序列可以用“标记-空白-空白-标记”来表示。在脉冲宽度调制（Pulse Width Modulation，PWM）记录技术中，标记的边沿表示位“1”（Skeda 等，1987）；因此，1001 序列可以用一定长度的单个标记来表示。更长的序列（如 10001）可以用更长的标记来表示，因此称为“脉冲宽度调制”。脉冲宽度调制技术可以实现比脉冲位置调制技术更高的线性记录密度。

图 25-5PWM 记录技术和 PPM 记录技术之间的比较。

注：相比而言，PWM 技术在实现时更加困难（需要更加复杂的通道），而且 PWM 技术的擦写过程对热效应更加敏感。因此，对于光驱提供商来说，PWM 的实现是一项具有挑战性的任务。

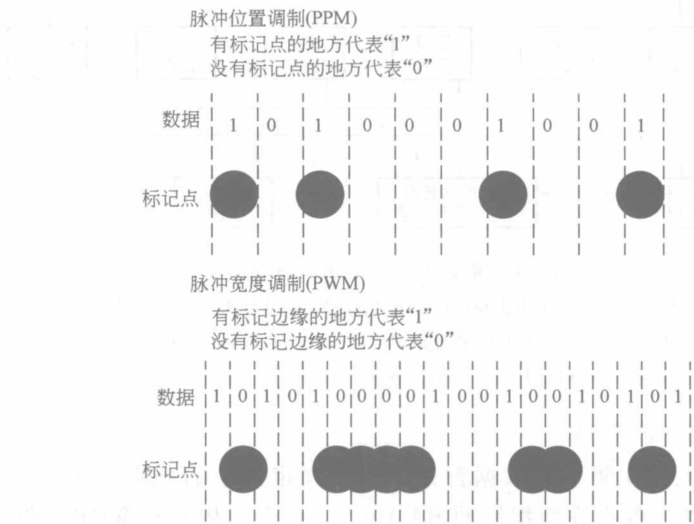


图 25-5 PPM 和 PWM 记录技术中标记间隔的示意图

注：PPM 记录技术是目前大多数光驱中使用的技术；PWM 记录技术可以提高约 50% 的容量，可用于可擦写光驱中。但是，相比 PPM 记录技术，PWM 记录技术要求更严格的容错能力

在读回过程中，光束从光盘上反射回来并入射到一个光电检波器（在磁-光驱动器中是两个检波器）。对于 WORM、相变和 CD-R 光盘来说，信号经过亮度

调制后，将通过检波器转换成电流，该电流在进行处理之前必须经过放大。WORM、相变和 CD-R 光盘上的标记和空白具有很高对比度，因此可以提供很高的信噪比。在磁-光驱动器中，从光盘上反射回来的读取信号不是进行亮度调制，而是进行“偏振”调制。因此，必须采用偏振光学和差分检波技术来将偏振调制转换成亮度调制（亮度调制在稍后的“磁-光记录”小节中将会介绍）。

为了从光电检波器产生的干扰信号中提取出 1 和 0，光驱采用了各种各样的技术（如均衡化），这些技术可以提高读写频率，而且还能在聚焦点之间提供更高的辨别力。使用模/数转换器时，模拟数据信号将被转换成通道数据流。在译码过程中，通道数据流又被重新转换成用户数据字节。数据在译码之前会添加时钟信号，这样可以消除调制模式。剩下的特殊字符从数据中提取出来之后，将被反馈到 ECC 排列缓冲器中，用于纠错（纠错能力最高可达 40B 的长度）。一旦数据从光盘中读取出来，这些数据将会保存在一个 RAM 缓冲器中，稍后将输出到连接 SCSI 和光驱的读取设备中。在了解了旋转光盘上数据的记录原理之后，接下来我们将详细介绍其中的各个专用术语，如用来阐述各种记录技术的记录物理学。

3. 相变记录技术

相变记录技术的原理是指某种材料可以处于多种亚稳态晶体相位，每种相位都具有各不相同的光学特性（如反射性）。在一定的热能作用下（由高功率激光束提供），这种材料就可以从一种亚稳态过渡到另一种亚稳态（Ovshinsky，1970；Takenaga 等，1983）。当提供的热能低于过渡临界值时，就不会发生状态转换。这样，低功率光束就可以用来读取记录信息，而且不会影响这些信息。

为了实现这类多种亚稳态，相变材料中混合了多种元素，如锗、碲和锑（$Ge_2Sb_2Te_5$）。相变材料既适合可重复擦写的记录技术，也适合一次性擦写的记录技术。在可擦除材料中，记录过程是通过激光束来软化材料并快速冷却将其冻结在一个无定形相位来实现的。快速冷却过程是关键，因此，材料散热性能的设计非常重要。

相变材料的可擦除性是通过一个退火过程实现的，即将材料加热到接近熔点并保持足够长的时间，以便在材料中进行再结晶，并擦除所有无定形标记。

由于相变材料是由多种材料混合而成的，因此，再结晶过程是很难实现的。熔化/退火过程加剧了相位偏析，从而降低了相变材料可循环使用的次数。早期的相变材料最多只能循环使用数千次，这也是当可擦写光驱诞生之后，很多公司淘汰相变材料而倾向于磁-光产品的主要原因之一。不过，目前相变材料的可循环使用次数已经大大提升了。

相比磁-光记录技术，相变记录技术的优势在于激光头的设计相对简单一些（因为不存在偏磁，而且需要更少的偏振光器件）；但相变记录技术的缺点包括

相变格式的标准化支持很少，而且生产这种光驱的公司很少。对于消费者来说，这就意味着会更加倾向于选择磁-光技术。

25.3 WORM 技术

事实上，最早的可擦写光驱是 WORM 光驱。尽管目前可擦写光驱在日常的存储需求中非常受欢迎，但是 WORM 技术在数据存储中仍然占据显著的位置，因为 WORM 技术具有恒久存档性能。

在商业产品中，包含了很多不同类型的一次性擦写技术。例如，可腐蚀的 WORM 光盘就是由碲基合金构成的；其中，数据的擦写过程是通过高功率激光在材料上灼烧一个小孔来实现（Kivits 等，1982）的。另一种 WORM 材料称为“织构化材料”，如虫眼模式，在实际应用中，常见的就是铂基薄膜。其中，数据擦写过程是通过将织构化材料熔化成一层平滑的薄膜从而改变其反射性来实现的。相变技术还提供了第三种 WORM 技术，该技术使用的材料是二氧化碲。在擦写过程中，无定形（模糊）材料通过激光束的热量转化成了晶格（光学）材料（Wrobel 等，1982）之后，这种材料的相位就再无法改变了。图 25-6 给出了相变 WORM 记录技术和烧蚀 WORM 记录技术之间的比较。

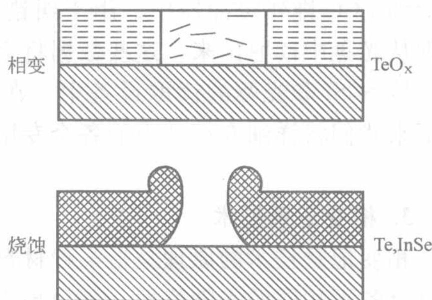


图 25-6 恒久 WORM 记录技术在可移动存储设备中提供了很高的数据安全性。在烧蚀 WORM 记录技术中，各个标记会被烧蚀到材料上，在相变 WORM 技术中，记录过程实际上是材料中相位变化的过程，该相位变化对应的是光反射性的变化

25.4 磁-光技术

在磁-光（Magneto-Optic，MO）驱动器中，数据记录过程是通过一个热磁过程实现的（Mayer，1958），该过程也称为“居里温度点写入”，因为该过程依赖于磁性材料的居里温度临界特性。在该过程中，光束聚焦点的能量会对记录材料进行加热，使其超过居里温度（约 200℃），高于该温度临界点时，材料中的磁畴就会对适度的外部磁场非常敏感（约 300 高斯（G））。外部磁场用来设置加热区域中磁化矢量的状态（磁化矢量描述了磁畴的偏振特性），状态为高表示位“1”，低表示位“0”。当材料的温度降低到居里温度点之下时，磁畴的方位就会固定下来。图 25-7 给出了磁-光记录、读取和擦除过程的示意图。磁-光记录过

程在任何区域都可以反复进行，而不会使材料的性能出现退化。对于可重复擦写的材料来说，这一特性非常重要。

在实际记录过程中，必须为记录设置一个明显的临界点，以确保一般环境下和读取过程中记录信息的稳定性。热磁记录过程是一个非常稳定的过程。除非材料加热到很高的温度（>100℃），磁-光记录材料的磁畴在数千高斯（G）的磁场下是不会受到影响的（相比而言，保存在磁性软盘中的信息，在 100G 的磁场下就会受到影响）。磁-光材料的矫顽力一直保持到温度非常接近于居里临界点温度时。当邻近居里临界温度时（约 200℃），磁-光材料的矫顽力就会以数量级的速度下降，直至磁畴结构变得模糊。

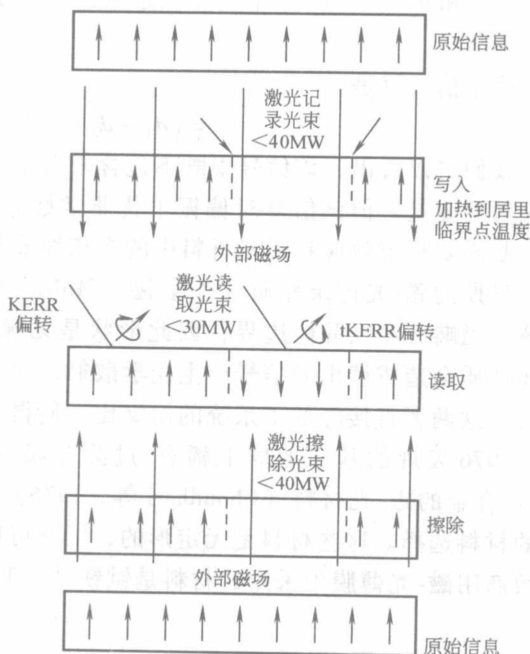


图 25-7　磁-光材料中的记录、读取和擦除过程示意图

当采用 2mW 功率的激光束时，就可以安全地读取出光盘上记录的信息；2mW 的功率已经可以为检波器提供足够的信号强度了，而且还不会影响到记录的信息，因为在该功率下介质的温度离居里临界点温度还很远。在信息读出过程中，记录信息位的磁性状态对低功率线性偏振读出光束的“极性 Kerr 效应”非常敏感。在该效应中，光束的偏振面会根据磁化矢量进行轻微偏转（0.5°）。旋转的方向会被读出检波器和通道检测到，该方向决定了位是“1”还是“0”。尽管轻微的 Kerr 偏转对较大的 DC 偏置上的信号调制不会产生很大的影响，但差分检测技术仍然可以实现可接受的信噪比（SNR）。

磁-光（MO）记录系统中输出信号来自于入射到某个检波器的光束减去入射到另一个检波器的光束时对应的信号。通过将偏振分光器放置在入射偏振的 45°方向，两个数据检波器就可以得到如下信号（Mansuripur，1982）：

$$d_1=\frac{I_0}{2}(\cos\theta_k/2)-\sin(\theta_k/2)^2\approx\frac{I_0}{2}(1-\theta_k)$$
$$d_2=\frac{I_0}{2}(\cos\theta_k/2)+\sin(\theta_k/2)^2\approx\frac{I_0}{2}(1+\theta_k) \tag{25-2}$$

式中，d_1和d_2是指检波器信号；I_0是指入射强度；$\theta_k/2$是指偏转角度，该角度通常很小。

读出信号就是

$$s=(d_1-d_2)/(d_1+d_2) \tag{25-3}$$

我们可以看出，该信号中既不包含来自激光的密集干扰，也不包含来自光盘的反射性变量。但该信号对偏振干扰非常敏感，这些偏振干扰是由偏振敏感衍射、基底双折射效应和磁光材料中的多相性或其他偏振敏感因素产生的。

早期的磁-光记录介质（如锰铋（MnBi）薄膜，Chen 等，1968）通常都是晶体。磁畴分布在晶体边界，因此形状是无规则的（Marchant，1990）。薄膜的晶体性质会造成读出光信号产生光学散射，而不规则的磁畴会在记录信号中产生干扰。这两者直接降低了系统的信噪比，使得多晶磁-光介质变得不实用。

1976 发现的基于地球上稀有/过渡金属（Rare Earth/Transition Metal，RE/TM）合金的磁-光材料（Choudhari 等，1976）为可擦写磁-光记录技术提供了实用的材料选择。这些材料是无定形的，因此可以实现可接受的信噪比。目前，大多数商用磁-光薄膜中采用的材料是铽铁钴（TbFeCO）。

25.5 可记录压缩盘

目前流行的可擦写可记录压缩盘（Compact Disc-Recordable，CD-R）看上去非常类似于印模 CD-ROM 光盘，在大多数 CD-ROM 播放器中都可以使用。

CD-音频和 CD-ROM 之所以成功，主要原因之一就是所有的生产商都遵守了严格的统一标准。这个标准是由 Philips 和 Sony 公司共同起草的，并在各种出版物中进行了详细阐述。其中，红皮书阐述了 CD-音频的标准，黄皮书阐述了 CD-ROM 标准，桔皮书阐述了 CD-R 标准。这些出版物描述了光盘必须满足的物理属性（如反射性、磁迹间距等等）和记录数据的分布（Bouwhuis 等，1985）。

在预录 CD-ROM 光盘中，数据信息是通过高反射性背景条件下的低反射性印模点来表示的。其中，光盘具有 70% 的反射性（通过铝膜层实现），而印模处具有 30% 的反射性（这些反射性规范在红皮书中进行了详细定义）。CD-ROM 光驱利用这种反射性上的差别来辨识光盘上的数据信息。为了与 CD-ROM 读取过程取得一致，CD-R 在记录数据信息时也必须利用这种反射性差别；为了实现这个目标，CD-R 上必须涂敷一层有机聚合物，该有机聚合物在激光束对其进行充分加热时可以永久改变其局部区域的反射性。图 25-8 给出了 CD-R 的结构示意图。

当有机染料聚合物的局部区域被激光束的聚焦点加热时，聚合物中的共价键就会被破坏，或者在局部区域形成复杂的折射率。这种折射率的变化直接造成了

光盘上记录材料反射性的变化。在商用光盘中，通常采用苯类有机聚合物，常见的如酞化青染料和多烷青色素。

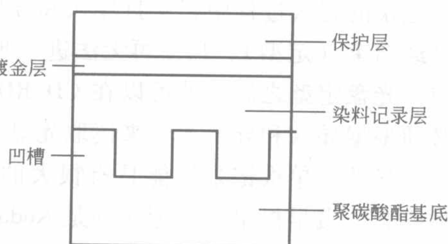


图 25-8 CD-R 的结构示意图

类似 CD-ROM 驱动器，CD-R 驱动器的性能也相对较低（与光驱和硬盘相比）。寻轨时间处于数百微秒的级别，而 4 倍转速光驱的最大数据传输率约为 600KB/s。寻轨时间很慢是因为 CD-R 驱动器是以“恒定线速度（Constant Line Velocity，CLV）”模式来驱动光盘旋转的，该模式在红皮书中进行了详细定义。恒定线速度意味着光盘的旋转速度会随其半径而变化，而激光头位于半径范围之内，以便线速度与半径保持不变。

但是，纯数据设备（如硬盘和 WORM 光驱）可以处理恒定角速度（Constant Angular Velocity，CAV）操作，在该操作过程中，线速度和数据传输率都会随着激光头辐射半径的增加而增加。恒定线速度操作的缺陷是会减慢访问速度。当光度头寻轨时，驱动引擎的速度必须根据激光头的辐射位置来进行调整；这个过程会耗费一定的时间，从而使寻轨过程延长 250 ~ 300ms。相反，恒定角速度设备与 WORM 光驱非常相似，其寻轨时间都在 40ms 级别。

CD-R 最初来源于音频 CD，因此部分参数、记录格式和性能特征都基于红皮书中的标准（该标准已经变成了一种束缚）。CD 格式（Bouwhuis 等，1985）不太适合面向模块的随即访问记录技术；CD 格式适合连续记录系统，如磁带录音机。

CD-R 的记录模式

为了准确掌握 CD-R 的属性和局限性，我们必须理解 CD-R 工作时的各种记录模式。对于固定模块结构设备来说，不存在记录模式问题，因为只有一种记录模式可选；但是在 CD-R 中，存在 4 种记录模式（Erlanger，1994）。

CD-R 驱动器中的 4 种记录模式分别为：

1）一次性光盘（或者单次记录）；

2）一次性寻轨；

3）多次记录；

4）递增性分组记录。

在一次性光盘记录模式中，无论记录数据是覆盖整个光盘还是其中一部分区域，在光盘上只能进行一次记录。单次记录光盘中的数据区域是由导入磁轨、数据域和导出磁轨构成的，导入磁轨中包含了目录表（Table Of Contents，TOC）。导入和导出磁轨对于 CD-ROM 驱动器和光盘之间的互操作来说非常重要。在单

次记录的写入过程中，一旦导入和导出区域被写入，系统就会认为光盘记录过程已经结束（定型），从而再无法进一步进行记录操作了，即使光盘中还有空白区域。光盘定型之后，就可以在CD-ROM播放器中进行播放了（CD-ROM播放器必须根据导入和导出磁轨来读取光盘）。

显然，单次记录性能具有很大的局限性，因此，我们引入了“多次”记录的概念。最早提出多次记录的是Kodak公司，他们希望在其照相CD产品中实现多次记录能力。在多次记录过程中，每一次记录都是根据导入和导出域来进行的。多次记录光盘可以在兼容多次记录光盘的CD-ROM驱动器中播放（假设每次记录都会在导入和导出域中定型）。但是，每次记录的导入和导出域都会占用很大的开销（约15MB）；因此，在一个容量为650MB的光盘上最终能实现的最大记录次数为45次。

如果不需要进行多次记录，用户可能会选择一次性寻轨记录技术。在这种记录技术中，每次记录过程都会对很多磁轨进行写入操作（各个磁轨代表了不同的写入模式）。整个光盘上可写入的最大磁轨数量为99个。但是，光盘在被CD-ROM驱动器读取之前必须被定型。

由于输入数据编码和展开方式的原因，在记录时，必须保持恒定的信息流。如果数据流出现中断，将会影响到整个记录的文件（不仅仅是MO或WORM驱动器中的一个扇区）。如果中断足够长，将会在光盘上形成一个空白区域，从而导致光盘中记录的数据变得无效。

上面这些问题可以通过一种称为“分组记录”的方法来缓解。在分组记录方法中，输入数据被分解成特定大小的数据分组（例如，128KB或1MB）。每个分组由1个连接块、4个导入块、1个数据域和2个导出块构成。导入和导出块用来描述各个分组，并提供重组空间，也就是说，当在不同的CD-R中记录相邻的数据分组时，如果没有实现很好的同步，导入和导出块就可以提供一定的重叠空间。

分组记录模式具有很多优点。首先，用于记录的分组数量没有限制（可与光盘的有效空间持平），而且，一次性寻轨、多次记录或一次性磁盘记录面临的限制都可以避免。其次，如果分组的大小小于驱动器缓冲区的大小（这也是真实情况），那么在记录时就不需要专用的硬盘了。一旦信息分组被发送到驱动器缓冲区中，那么计算机就可以进行脱机工作，而CD-R驱动器自行完成记录操作。

基于分组记录技术的这些优点，CD-R技术在过去变得越来越灵活，并因此成为了通用可移动存储设备。CD-R还可以用来备份保存小型文件。但是，CD-R与CD-ROM播放器之间的互操作存在一个问题，即如果CD-R在进行分组写入时，在每个分组开始处的连接块中产生了硬错误，那么CD-ROM播放器就不

能读取该 CD-R。分组写入的 CD-R 可以在 CD-R 驱动器和支持分组的 CD-ROM 驱动器上读取。

25.6 光学磁盘系统

1. 磁盘

光学存储系统中的一个重要组成部分就是介质磁盘。事实上，光学存储系统最吸引人的地方就是存储介质的可移动性，如软盘。大多数有效可写入介质的磁盘都是 3.5 或 5.25in。例如，图 25-9 给出了 5.25in 磁盘的结构示意图。磁盘上包含很多传感器孔，用来方便驱动器对其介质类型进行识别。介质上的磁轨是以螺旋的形式排列的。第一块数据记录在最里面的磁轨上（接近中心）。可记录 CD 介质需要磁盘盒保护，在 CD-R/CD-ROM 驱动器中播放时放在一个托盘上或放在专用的盘槽中（如音频 CD 播放器）。

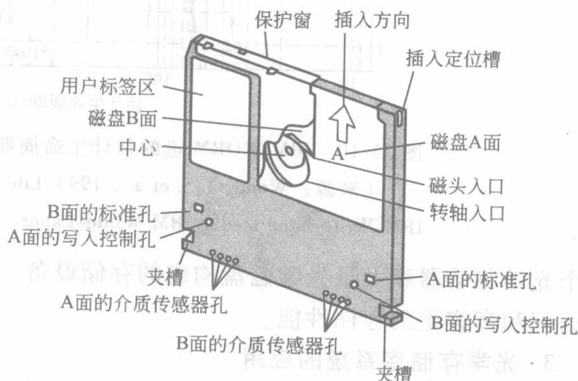


图 25-9　5.25in 磁盘的结构示意图

注：磁盘中的各种孔隙可以被驱动器识别，用来提供磁盘的相关信息

磁盘介质具有很长的保存期限和存档期限（没有特殊的和昂贵的环境控制），而且可以用来保存自动读盘机和用户操作的精确信息。磁-光 WORM 介质非常稳定，因此保存在该介质中的数据信息具有很高的可靠性（Okino，1987）。图 25-10 给出了 IBM WORM 介质的估计生命周期对数正态分布图。在该图中，描述了在温度/湿度条件为 30℃/80% 时，97.5% 介质表面分别用于保存和存档目的的保存时间，分别超过 36 年和 510 年。保存期限是指数据可进行有效写入的保存时间；而存档期限是指数据可进行有效读取的保存时间。

光学介质可能是最稳定的数字存储技术，因为磁性驱动器容易产生磁头划痕、磁带退化或透印（在透印中，如果磁带产生轻微的损伤，介质上某一层的信息会转移到其他层面上去），而且纸带或缩微胶片也会随时间逐渐退化（Rothenberg，1995）。

2. 自动光学存储系统

光学驱动器可以扩展到自动存储系统中，自动存储系统实质上就是由一个或

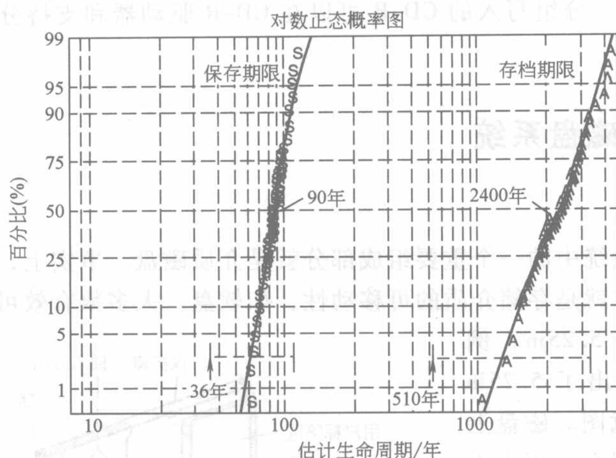


图 25-10 IBM WORM 盘的估计生命周期对数正态分布图

（来源：Wong, J. S. et a.，1993. Life expectancy of IBM Write-once media. IBM White-Paper, IBM Publication.）

多个光学驱动器和大量光学磁盘构成的存储设备。光学存储库可以提供在线、直接访问和高容量的存储性能。

3. 光学存储库系统的应用

光学存储库系统非常适合计算机环境，如客户-服务器系统、对等局域网或框架系统。在这些环境中，有非常明显的分级存储体系，该分级存储体系的基础是各种类型数据的成本/访问性能比较。图 25-11 给出了分级存储体系的结构示意图，其形状像一个金字塔。金字塔结构中的最高层是性能最好、成本最高的存储设备类型；最底层的是最廉价（成本/MB）和性能最差的存储设备类型。光学存储库是其中的重要组成部分，因为它提供的存储性能接近磁性存储设备，但其成本却只与磁带相当。

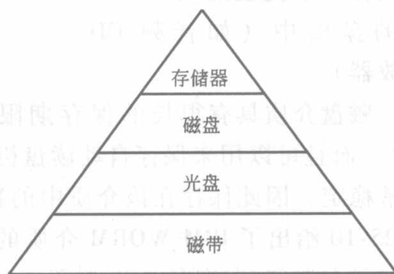


图 25-11 分级存储体系

光学存储库中包含了一种磁盘转移机制，称为“自动变换器”。自动变换器可以在输入输出槽（光盘通过该槽插入到存储库中）、驱动器（磁盘的读写位置）和磁盘存储单元之间转移光盘。

复杂的存储系统可以提供分级存储管理功能，在分级管理功能中，数据根据访问需求自动在金字塔的各层之间迁移。例如，电话公司可以在磁性存储器中保存当前的账单信息，而 1 个月之前的账单信息可以保存在光学存储库中，6 个月

之前的账单信息可以保存在磁带中。系统访问不频繁的数据不必保存在成本较高的存储系统中，如半导体存储器或高性能磁性设备。

自动光学存储最理想的应用之一就是文档成像技术。在目前的应用中，文档成像直接涉及到纸张管理问题。在过去一个世纪的时间里，我们保存和管理纸张文档的方式都没有发生很大的变化，而且以前在我们的工作场所都没有广泛使用计算机（很多公司都是采用仓库来存放文件柜和文档盒）。现在，信息的保存方式可以用金字塔模型来描述（见图25-12），在金字塔模型中，纸张几乎囊括了所有的文档存储形式。

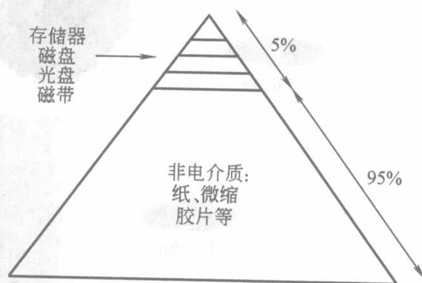


图25-12　文档保存金字塔结构

目前，世界上大多数的文档都是以纸张的形式保存的，只有5%的文档是以电子版的形式保存的。

文档成像技术直接对我们处理和存储文档的方式进行了重新设计。在文档成像技术中，文档首先被扫描到计算机中并保存为计算机可读取的文档。将文档保存为计算机格式具有很多好处，包括可以快速访问信息，或者对保存的文档调用时，只需要输入几个搜索文字就可以了。但是，文档成像对存储空间要求很高，因此，就需要高容量、低成本的存储技术，以满足快速随机访问的需要。光学存储是目前认为最佳的文档成像存储技术。光学存储库可以根据大容量需求来进行调节，以存储大量的文档图像（文字或照片），而且光驱可以任意对这些存储的图像进行访问。图25-13介绍了单个容量为1.3GB的光盘与相同存储容量设备之间的数量对比。

4. 光学存储系统技术的未来与发展

光学驱动器和存储库将朝两个方向发展（类似于任何高技术产品）：第一个就是渐增型改进过程，在该过程中，存储技术的质量和功能将会不断提高；第二种是更加明显的改进过程，在该过程中，光盘容量和驱动性能将会以每隔几年一个台阶的速度提高。

对于存储库来说，改进的方面主要集中在自动定位机制的速度以及数据和存储设备管理算法的实现上。存储库系统中的大多数改进主要来自于驱动器自身的改进。

对于光学驱动器来说，渐增型改进过程主要集中在核心技术单元上，如：激光器、介质、记录通道和光机。激光器可以改进的方面包括光束质量（降低波像差和散光）、使用寿命（提供更长的保险期）以及功率（进一步加快光盘的旋转速度）；介质可以改进的方面包括基底（降低倾斜和双折射的产生）、活性层

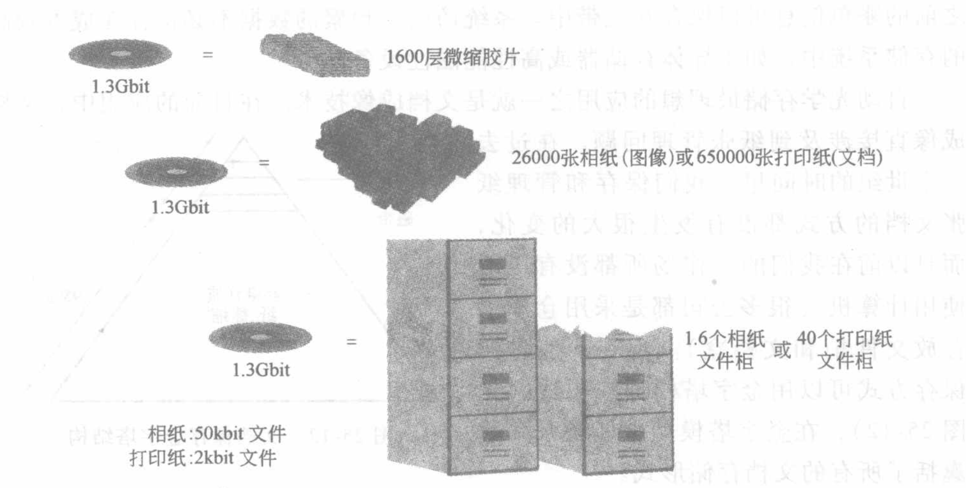


图 25-13　单个容量为 1.3GB 的光盘与相同存储容量的微缩胶片、纸张和文件柜之间的数量对比

(提高感光性）和钝化层（提高适用期限）；光驱和传动装置可以改进的方面包括伺服系统（提高定位性能)、降低干扰、减小光学器件尺寸以及减轻传动装置的质量（提高寻轨速度)；记录通道和电子技术可以改进的方面包括电子集成度的提高、电子干扰的降低、低功率电子元器件的使用以及更好的信号处理和ECC（提高数据可靠性)。其中，最重要的改进方向之一就是不断降低驱动器和介质的成本，这是提高光学驱动器在市场上畅销度的必要步骤。

在基本的改进方面，光学驱动器有很多技术可以选择。未来光学驱动器的发展主要集中在两个方向上：提高容量；提高性能规范（如数据速率和寻轨时间)。图 25-14 描述了可以实现上面两个目标的技术。

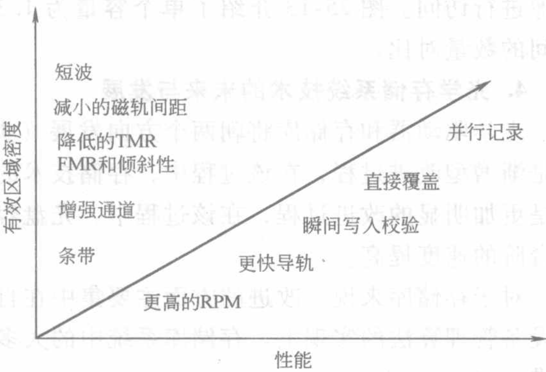


图 25-14　光学驱动器性能和容量发展的技术方向

性能上的发展包括提高光盘旋转时的分辨率，从而实现更高的数据传输速率，同时从根本上提高寻轨速度，以便快速访问数据，以及提高直接覆盖和瞬间校验技术，从而降低记录时磁头来回运动的次数。最后，并行记录技术的使用（多元激光或特殊光学器件）将可以在很大程度上提高数据传输速率。

在光盘产品中，要实现更高的容量只有通过提高表面的数据密度才能实现，因为光盘的尺寸在目前的标准尺寸条件下是无法提高的。在提高光学驱动器的存储容量方面，有很多种技术可供选择，包括：

1）采用短波激光或超级分辨率技术来实现更小的聚焦点尺寸；

2）将数据磁轨靠的更近，以实现更高的表面密度；

3）降低对聚焦重合失调（Focus MisRegisteration，FMR）（如散焦）和寻轨重合失调（Tracking MisRegisteration，TMR）（如当聚焦点没有集中到磁轨上）的灵敏性；

4）降低对介质倾斜的灵敏性。

最终，读取通道方面的改进（如利用部分响应最大相似性（Partial Response Maximum Likelihood，PRML）技术）可以在高密度记录中存在干扰时使标记可以更加容易被识别。

感谢

在此，作者要感谢两位 IBM Tucson 公司的同事：Blair 和 Alan Fennema；他们分别为读取通道章节和伺服系统中的擦写章节给出了意见和建议。

名词解释

可擦除压缩磁盘/可记录压缩磁盘（CD-E/CD-R）：定义了压缩盘的可写入版本；CD-E 采用的是可擦写介质，而 CD-R 采用的是一次性写入介质。

数据块大小：记录在光盘上的数据会被格式化成最小的数据块尺寸。标准尺寸是 512B，该标准为信息定义了一个单矢量。很多 DOS/Windows 程序中都支持这个标准的数据块尺寸。如果数据块的尺寸可以更大的话，那么存储系统的效率就会变得更高。数据块的另一个飞跃式尺寸发展就是 1024B，通常用在 UNIX 应用程序中。

设备驱动：这是一种软件，可以使主机与光学驱动器实现互操作。如果没有设备驱动软件，我们就无法将光学驱动器连接到 PC 上，从而无法工作。随着光学驱动器的不断发展，设备驱动也将会集成在操作系统中。例如，CD-ROM 驱动器的驱动软件就是嵌入在操作系统中的。

纠错控制（ECC）：添加在原始数据中，用来进行错误检测和纠正的代码。纠错码包含很多种类，例如，压缩磁盘采用交叉 Reed-Solomon 纠错码。

磁-光介质：一种光学记录材料，在该类型材料中，标记是通过热磁过程进行记录的。也就是说，材料被加热，直到磁畴可以在模式磁场的作用下发生变

化。该材料是可写入的材料。

光学读盘机：在概念上与传统的自动播放机非常相似。在播放机中包含了很多磁盘，可以进行随机读写。

相变介质：一种光学记录材料，由合金构成，合金中包含两种具有不同光学特性的亚稳态相位。

脉冲位置调制（PPM）：一种记录技术。在该技术中，磁盘上的标记表示二进制位"1"，空白表示二进制位"0"。1001 的序列可以表示成"标记-空白-空白-标记"。

脉冲宽度调制（PWM）：一种记录技术。在该技术中，标记的边缘用来表示"1"，而没有边缘的地方表示"0"。因此，1001 的序列就可以通过一个标记来表示。

寻轨时间/访问时间：这两个概念经常混淆使用，这是不对的。按照惯例，寻轨时间是指传动装置搜寻超过 1/3 距离（光盘上从内部数据带到外部数据带的距离）所消耗的时间；而访问时间是指寻轨时间加上传动装置的定位反应时间和光盘旋转到合适速度所消耗的时间。访问时间真正描述了我们访问数据的速度。

伺服系统：一种机制，通过反馈和控制来保持激光束聚焦在光盘的磁轨上。在转速为 4000r/min 的光盘上，这是一个复杂的过程。

寻轨：光盘上"光笔"（即激光束）保持在数据磁轨中心的方式。

参考文献

[1] Arimoto, A. et al. 1986. Optimum conditions for high frequency noise reduction method in optical videodisk players. *Appl. Opt.* 25 (9): 1.

[2] Asthana, P. 1994. A long road to overnight success. *IEEE Spectrum* 31 (10): 60.

[3] Bouwhuis, G., Braat, J., Huijser, A., Pasman, J., van Rosemalem, G., and Schouhamer Immink, K. 1985. *Principles of Optical Disk Systems.* Adam Hilger, Brustol, England, UK.

[4] Braat, J. and Bouwhuis, G. 1978. Position sensing in video disk read-out. *Appl. Opt.* 17: 2013.

[5] Chen, D., Ready, J., and Bernal, G. 1968. MnBi thin films: Physical properties and memory applications. *J. Appl. Phys.* 39: 3916.

[6] Choudhari, P., Cuomo, J., Gambino, R., and McGuire, T. 1976. U. S. Patent #3, 949, 387.

[7] Earman, A. 1982. Optical focus servo for optical disk mass data storage system application. *SPIE Proceedings* 329: 89.

[8] Erlanger, L. 1994. Roll your own CD. *PC Mag.* (May 17): 155.

[9] Golomb, S. 1986. Optical disk error correction. *BYTE* (May): 203.

[10] Goodman, J. 1968. *Introduction to Fourier Optics.* McGraw-Hill, San Francisco.

[11] Inoue, A. and Muramatsu, E. 1994. Wavelength dependency of CD-R. *Proceedings of the Optical Data Storage Conference.* Optical Society of American, May, p. 6.

[12] Kivits, P., de Bont, R., Jacobs, B., and Zalm, P. 1982. The hole formation process in tellerium layers for optical data storage. *Thin Solid Films* 87: 215.

[13] Mansuripur, M., Connell, G., and Goodman, J. W. 1982. Signal and noise in magnet-optical readout. *J. Appl. Phys.* 53: 4485.

[14] Mansuripur, M. 1987. Analysis of astigmatic focusing and push-pull tracking error signals in magnet-optical disk systems. *Appl. Opt.* 26: 3981.

[15] Marchant, A. 1990. *Optical Recording*. Addison-Wesley, Reading, MA.

[16] Mayer, L. 1958. Curie point writing on magnetic films. *J. Appl. Phys.* 29: 1003.

[17] Okino, Y. 1987. Reliability test of write-once optical disk. *Japanese J. Appl. Phys.* 26.

[18] Ovshinsky, S. 1970. Method and apparatus for storing and retrieving information. U. S. Patent #3, 530, 441.

[19] Rothenberg, J. 1995. Ensuring the longevity of digital documents. *Sci. Am.* (Jan).

[20] Sukeda, H., Ojima, M., Takahashi, M., and Maeda, T. 1987. High density magneto-optic disk using highly controlled pit-edge recording. *Japanese J. Appl. Phys.* 26: 243.

[21] Takenaga, M. et al. 1983. New optical erasable medium using tellurium suboxide thin film. *SPIE Proc.* 420: 173.

[22] Tarzaiski, R. 1983. Selection of 3f (1, 7) code for improving packaging density on optical disk recorders. *SPIE Proc.* 421: 113.

[23] Treves, D. and Bloomberg, D. 1986. Signal, noise, and codes in optical memories. *Optical Eng.* 25: 881.

[24] Wrobel, J., Marchant, A., and Howe, D. 1982. Laser marking of thin organic films. *Appl. Phys. Lett.* 40: 928.

备注

关于光盘系统的介绍，有很多优秀的书籍可供参考。其中，有 Alan Marchant 的《*Optical Recording*》(Addison-Wesley, Reading, MA, 1990)，该书介绍了各种类型的记录技术以及光驱的基本机能。关于光学磁盘驱动器及其光学机械原理更详细的研究，读者可以参考 G. Bouwhuis、J. Braat、A. Huijiser、J. Pasman、G. van Rosemalen 和 K. Schouhamer Immink 的《*Principles of Optical Disc Systems*》(Adam Hilger Ltd., Bristol, England, 1985)。关于磁-光记录技术的研究，读者可以参考 Masud Mansuripur 的《*The Physical Properties of Magneto-Optical Recording*》(Cambridge University Press, London, 1994)。

关于光学存储领域的最新发展，建议读者留意"光学存储器国际座谈会（ISOM）"或"光学数据存储（ODS）会议"（由 IEEE 或美国光学协会赞助）。

关于光学存储系统及其应用的发展，读者可以参考行业杂志：《*Computer Technology Review*》。另外，《*Imaging Magazine*》中通常也包含了很多关于文档成像的光学存储库应用以及定期评论商业光学产品的文章。

第 26 章　纠 错 技 术

Fabrizio Pollara

26.1　背景知识

数字信号在信道上传输时会产生损伤，如噪声干扰、失真等，因此信号在到达用户时就必然会出现错误，从而产生衰减。保存在具有缺陷的磁性介质或其他介质中的数字数据，也会出现类似的情形。对于数字通信链接和数据存储来说，误码率是一项非常重要的设计规范。通常，误码率必须保持在预定值之下，在不同的应用中该预定值也各不相同。在原始消息上添加冗余信息的纠错技术可以用来控制该误码率。最常见的例子就是重发相同的消息（即使这种方式的效率不高）。本章中我们将详细介绍几种高效的纠错技术。

1948 年，Claude Shannon 的经典著作宣布了通信技术革命的到来。在 Shannon 的著作中，他详细阐述了消息在发送到信道之前，对消息进行编码的最佳方案（例如引入长度可控制的冗余数据），这是保证通信过程可靠性必不可少的步骤。Shannon 并没有指明具体的编码过程如何进行，他只是在数学上证明了高效编码方案的存在。自 1948 年以来，很多研究人员都通过直接和实际的编码方案证明了 Shannon 理论的正确性，这些编码方案现在已经应用到了现代数字通信系统中。卫星通信系统、高性能军事通信系统、计算机通信网络、高速调制解调器和光盘记录和播放系统等，所有这些系统都依赖于复杂的编码方案来提高它们的性能。

Shannon 指出，对于每个通信信道，我们都可以为其定义一个信道容量 C；而且总是存在一种纠错编码，使得信息可以以低于 C 的速率在噪声信道上传输，同时误码率可以任意小。事实上，Shannon 理论中一个很重要的结论就是没有必要构建过于理想的信道，只需使用比较经济的编码方案就足够了。在 Shannon 理论诞生之前，我们一直认为克服噪声信道的惟一方法就是使用功率更高的发射器或者构建更大的天线。

通信系统将数据源和数据用户通过信道连接到了一起。微波连接、同轴电缆、电话线路以及磁带都是信道的一种形式。编译码数字通信系统（或数据存储系统）中包含一个用来生成适当冗余数据流的设备（编码器）和一个对应的利用这些冗余数据来纠正信道码字错误的设备（译码器），如图 26-1 所示。编

码器中包含大量的信息并生成了很多信道符号，称为“码字”。图 26-1 中的信道包含了在物理介质上传输数字信息时所需的所有必要设备，包括一个调制器（用来将码字中的每个符号转换成对应的模拟符号（波形））和一个解调器（用来将接收到的每个信道输出信号转换成码字）。每个被解调后的符号都是发送符号的一个最佳估计值，但由于信道干扰，解调器也会产生一些误码。



图 26-1　编码通信系统组成图

26.2　引言

高效纠错方案的基本概念在下面的汉明（Hamming）码中可以得到很好的解释。Hamming 码是一种只能校正一位错误位的纠错码，属于第一类实用的编码技术，是由 Hamming 在 1948 年发明的。

在单位纠错码（Hamming 码）中，发送的数据由 4bit 块构成，分别为 $\{d_1, d_2, d_3, d_4\}$，3 个冗余位 $\{p_1, p_2, p_3\}$（奇偶校验位）是根据数据块中的位数计算得到的，并通过编码器附加到该数据块中。如果 d_2，d_3和 d_4中位“1”的数目为偶数，那么奇偶校验位p_1就会被设置为 0；否则，p_1就会被设置为 1。类似地，奇偶校验位p_2根据 d_1，d_3和 d_4中位“1”数目的奇偶性被设置为 0 或 1（即奇或偶）。最后，p_3是根据 d_1，d_2和 d_4 中位“1”数目的奇偶性来进行设置的。举例来说，如果数据流为 {1011}，那么奇偶位就是 {010}。之后，数据位和奇偶位合并在一起被发送到信道上，或保存在一个存储器中，这些一起称为“码字”。

当码字被恢复之后，其中可能会包含一些误码，译码器就会对其进行检查来核对恢复后的数据流是否仍然满足奇偶校验规则。译码器对合法的奇偶校验输出 1，对错误的奇偶校验输出 0，这样，最后就可以得到 3 位，称为“伴随式”。例如，如果接收到的为 {1001010}，那么第一个伴随式数据位就会被设置为 1，因为接收到的数据中 $d_2 = 0$，$d_3 = 0$，$d_4 = 1$，它们的奇偶性为奇数，但第一个奇偶校验位为 0。类似地，第二个和第三个伴随式数据位分别设置为 1 和 0，从而得到伴随式 {110}。稍后，该伴随式就会根据伴随式和误码位置之间的一对一对应关系来确定误码的位置，如图 26-2 所示。在该例子中，伴随式 {110} 表明误码出现在 d_3 位置，即第 3 个数据位。然后译码器就会将 {1001010} 变成 {1011010}，并将正确的数据 {1011} 发送给用户。这种纠错码的奇偶校验方法可以成功定位 4bit 数据块中的任何单个误码，该 4bit 数据块中添加了 3 个奇偶校

验位。如果误码位数超过1个，那么Hamming纠错码就无效了。

伴随式	000	001	010	011	100	101	110	111
误码位置	–	p_3	p_2	d_1	p_1	d_2	d_3	d_4

图 26-2 伴随式表

目前，在很多应用中都包含了各种有效的编译码方案，可以用来纠正更多的误码。从实用的角度来看，所有编译码方案存在的本质局限性不是指缺少好的编码方法，而是指译码器的复杂程度（和成本）。因此，编码和译码方案的设计就非常容易实现了。

纠错编码的优势

在编码过程中插入到数据中的冗余数据数量通常称为“编码率 R”。R 是数据块中信息位数 k 与码字中总的传输符号数 n 的比值（$n = k +$ 冗余符号数；即 $n > k$ 或者 $R = k/n < 1$）。

我们假设要传输的数据速率为 R_b bit/s，代码冗余量迫使我们必须将发送符号速率 R_s 提高到 $R_s = R_b/R > R_b$。如果每个信息位的发送能量 E_b 是固定的，那么每个发送符号的接收能量就会从 E_b 降低到 $E_s = R \times E_b$。如果我们不进行任何特殊的译码操作，那么误码率（BER）将会比没有使用编码时的原始值高很多。不过，如果我们使用冗余位来校正误码并选择合适的编译码方案的话，译码器输出端的BER就会比无编码系统的原始值低很多。BER的这种提高是在与无编码系统相同 E_b 和 R_b 条件下实现的，这就是为什么在给定 E_b 和 R_b 条件下需要采用编码来降低BER的原因。另外，如果给定的条件是BER和 R_b，那么编码过程可以有效降低 E_b；或者，如果给定的条件是BER和 E_b，那么编码过程可以有效提高 R_b（吞吐量）。在给定的编码系统中，吞吐量提高或传输功率下降的比例称为“编码增益”。

26.3 纠错码的发展

根据Shannon理论，Hamming纠错码与其他更长的纠错码比较起来就显得没有优势了。继Shannon之后，Bose和Ray-Chaudhuri（1960）以及Hocquenghem（1959）发现了可以纠正多个误码的一类多位纠错码（BCH码），Reed和Solomon（1960）发现了非二进制信道上的一类相关纠错码。所有这些纠错码的基础都是有限域（Galois域）的代数性质。

真正完全不同的高效纠错码是20世纪50年代末发现的“卷积码”，卷积码基于移位寄存器生成的二进制序列特性。直到1967年发现了有效的译码规律后，卷积码才开始真正得到广泛应用。直到20世纪90年代初，代数“块状码”与卷积码的级联都是效率最高的编码应用方案。

20世纪70年代，纠错码技术的主要发展就是纠错码系列的发现，这些纠错

码是渐近的，也就是说，编码消息块的大小可以是任意的。效率更高的渐近纠错码（基于代数-几何特性）直到20世纪80年代末才被发现。

20世纪80年代初，人们发现了在有限带宽信道上可以进行高效率传输的纠错码，该纠错码通过选择合适的调制方案来合并编码过程；也就是说，在实际中使用各种波形来传输编码符号。这些纠错码在任何商业数据通信系统的设计中都具有非常重要的实际意义。

最近，一种新的纠错码（Turbo 码）填补了 Shannon 理论中关于编码极限的空缺。

26.4　纠错码系列

纠错码可划分为两种不同的类型：块状码和卷积码。块状码的译码器首先将连续的信息流分解成各个长度为 k bit 的分段或者块，然后根据使用的特定纠错码来单独对每个代码块进行处理。对于每个可能的信息位，代码块都与一个 n 元信道符号相互关联，其中，$n>k$；数量 n 称为“代码长度”或“块长度”。代码的长度在定义时可以比二进制字母表更长，这样代码中每个信息采样值就不是1bit了，而是来自字母表中的一个具有 q 项特性的符号了。

其他类型的纠错码称为“卷积码”，卷积码不会将信息序列分解成各个独立的代码块。卷积码的译码器连续处理信息并将每个较长的信息序列与一个包含很多符号的代码序列相关联起来。

块状码和卷积码具有相似的纠错能力和相同的基本局限性。另外，Shannon 基础理论在这两种纠错码中都适用。

哪种纠错码是最佳的纠错码？

块状码可以通过3个参数来进行评价：块长度 n、信息长度 k 和最短距离 d。最短距离是指两种相似的码字之间不同位的数量。

两个长度为 n 的 q-进制序列 x 和 y 之间的汉明距离 $d(x, y)$ 是指它们之间不同位的数量。一个纠错码的最短距离是指一对码字的汉明距离，这对码字具有最小汉明距离。纠错码之间更加精确的比较可以从纠错码的“重量分布”来分析，重量分布是指每个码字与给定参考码字之间的一列汉明距离。后面我们将会发现，对于线性纠错码来说，参考码字的选择是无关紧要的。

假设传输了一个码字，而且信道中产生单个误码。那么，接收到的码字与发送码字之间的汉明距离就是1。如果该码字与其他每个码字之间的距离都大于1，那么如果最靠近接收码字的码字实际上就是发送的码字，译码器将会即时校正该误码。通常，如果产生 t 个误码，而且接收到的码字与其他所有码字之间的距离都超过 d，那么如果最靠近接收码字的码字实际上就是发送的码字，译码器将会

对误码进行校正。这种方法只适用于 $d \geqslant 2t+1$ 的情况，如图 26-3 所示；在该图中，每个半径为 t 的球都是由距离为 t 的 n 元码字 C 构成，我们假设该码字是球的中心。半径为 t 的不相交的球可以描绘成各个单独的码字。每个球中接收到的所有码字都当作球中心的码字进行译码。如果产生了 t 个或更少个误码，那么接收到的码字将始终处于相应的球内，而且译码得到的结果也是正确的。那些误码数超过 t 的接收码字将处于另一个译码球内，因此，译码得到的结果就是错误的。其他误码数超过 t 的接收码字位于各个译码球之间的空隙中，这种情况可以通过两种方式来处理：利用“不完全译码”对那些位于某个译码球内的接收码字进行译码，而其他接收码字的误码数超过了允许的误码数，译码器将会认为它们无法识别（无法校正误码模式）；利用“完全译码”通过选择最近的码字来对每一个接收到的码字进行译码（无论是否位于译码球内）。当产生的误码数超过 t 时，完全译码器通常无法得到正确的译码结果，但偶然会发现正确的码字。当需要对消息进行最佳推测而不是估计时，可以使用完全译码器。

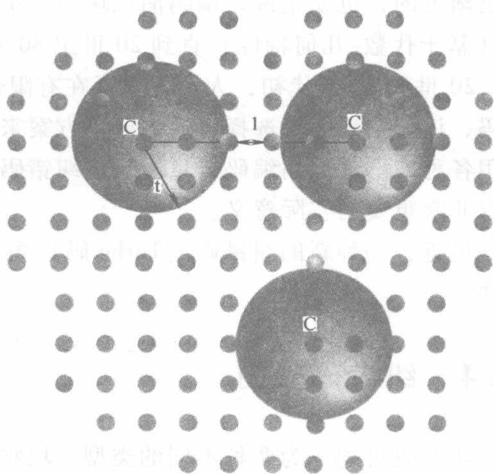


图 26-3　纠错能力为 t 的纠错码几何表示法（$t=2$）

26.5 线性块状码

块状编码过程可以看做是一个表格查找过程；在该过程中，每 $M=q^k$（q 是信息符号的字母表大小）个码字 x_1，x_2，…，x_m，…，x_M 保存在内存的一个 n 阶寄存器中；无论消息 $\boldsymbol{u}_m=(u_{m1}, \cdots, u_{mk})$ 何时发送，对应的信号矢量 $\boldsymbol{x}_m=(x_{m1}, \cdots, x_{mn})$ 都将会被从存储器中读出并作为译码器的输出值。我们主要讨论的是二进制字母表，也就是 $q=2$；因此，一开始，我们令所有的 u_{mi} 为 0 或者 1。

二进制数据的线性编码过程可以描述如下：

$$\begin{aligned}
x_{m1} &= u_{m1}g_{11} \oplus u_{m2}g_{21} \oplus \cdots \oplus u_{mk}g_{k1} \\
x_{m2} &= u_{m1}g_{12} \oplus u_{m2}g_{22} \oplus \cdots \oplus u_{mk}g_{k2} \\
&\vdots \\
x_{mn} &= u_{m1}g_{1n} \oplus u_{m2}g_{2n} \oplus \cdots \oplus u_{mk}g_{kn}
\end{aligned} \tag{26-1}$$

式中，⊕表示模2加法；对于所有的 i，j 来说，$g_{ij} \in \{0, 1\}$。$u_{mi}g_{ij}$ 是一般的乘法运算，因此，只有当 $g_{in}=1$ 时，x_{mj} 的运算才需要 u_{mi}。下面的矩阵称为线性码的“生成矩阵”，其中 $\boldsymbol{g}_i$ 是行矢量。

$$\boldsymbol{G}=\begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ & \vdots & & \\ g_{k1} & g_{k2} & \cdots & g_{kn} \end{bmatrix}=\begin{bmatrix} \boldsymbol{g}_1 \\ \boldsymbol{g}_2 \\ \vdots \\ \boldsymbol{g}_k \end{bmatrix} \tag{26-2}$$

因此，式（26-1）可以表示成矩阵形式：$\boldsymbol{x}_m=\boldsymbol{u}_m\boldsymbol{G}=\sum_{i=1}^{k} u_{mi}\boldsymbol{g}_i$；其中，$\Sigma$ 表示模2加法，$\boldsymbol{u}_m$ 和 $\boldsymbol{x}_m$ 都是二进制行矢量。

线性码的一个重要特性就是码字 $\boldsymbol{x}_m$ 和 $\boldsymbol{x}_l$ 的逐项模2求和结果仍然是一个码字。这样就会产生一个有趣的现象，即对于所有码字来说，给定码字与（$M-1$）个其他码字之间的汉明距离都是相同的。

我们可以发现，生成矩阵 $\boldsymbol{G}$ 可以重新排列，以便每个码字的前 k 个分量与消息 $\boldsymbol{u}_m$ 的前 k 个分量完全相同，而剩余的 $n-k$ 个分量都是奇偶校验符号，如同汉明纠错码例子中的奇偶校验符号。该例子中的编码称为“系统码”。

1. 线性块状码的译码过程

假设消息 $\boldsymbol{u}=u_1$，…，u_k 被编码成码字 $\boldsymbol{x}=x_1$，…，x_n，该码字稍后将会发送到信道中。由于存在信道干扰，因此接收到的矢量 $\boldsymbol{y}=y_1$，…，y_n 可能会与码字 $\boldsymbol{x}$ 有所不同。我们定义误码矢量为 $\boldsymbol{e}=\boldsymbol{y}\oplus\boldsymbol{x}=(e_1, \cdots, e_n)$。

译码器必须根据 $\boldsymbol{y}$ 来确定发送的是哪一个消息 $\boldsymbol{u}$ 或码字 $\boldsymbol{x}$。由于 $\boldsymbol{x}=\boldsymbol{y}\oplus\boldsymbol{e}$，因此如果译码器知道了 $\boldsymbol{e}$，就可以知道是哪一个消息 $\boldsymbol{u}$ 或码字 $\boldsymbol{x}$ 了。但是，译码器根本无法确定真正的 $\boldsymbol{e}$。因此，必须采取一定的策略来选择最相似的误码矢量 $\boldsymbol{e}$，同时假设 $\boldsymbol{y}$ 已经接收到。即假设所有的码字都是非常相似的，这种策略就是最佳的选择，因为它可以使译码器出现错误的概率降到最低，因此称为“最大似然译码”。

下面，我们详细解释系统线性码译码过程中的表格查询技术，该技术在汉明纠错码中非常普遍。我们首先定义一个 $n\times(n-k)$ 阶矩阵 $\boldsymbol{H}^{\mathrm{T}}$，如下所示：

$$\boldsymbol{H}^{\mathrm{T}}=\begin{bmatrix} g_{1,k+1}\cdots g_{1,n} \\ g_{2,k+1}\cdots g_{2,n} \\ \vdots \\ g_{k,k+1}\cdots g_{k,n} \\ 1\,0\,0\cdots\cdots 0 \\ 0\,1\,0\cdots\cdots 0 \\ \vdots \\ 0\,0\,0\cdots\cdots 1 \end{bmatrix} \tag{26-3}$$

矩阵 $\boldsymbol{H}$ 称为“奇偶校验矩阵”。任何码矢量乘以矩阵 $\boldsymbol{H}^{\mathrm{T}}$ 后可以得到零矢量 $\boldsymbol{x}\boldsymbol{H}^{\mathrm{T}}=0$。现在假设将任何接收到的矢量 $\boldsymbol{y}$ 右乘矩阵 $\boldsymbol{H}^{\mathrm{T}}$，得到的 $(n-k)$ 维二进制矢量称为接收矢量的“伴随式 $\boldsymbol{s}$”，公式为 $\boldsymbol{s}=\boldsymbol{y}\boldsymbol{H}^{\mathrm{T}}$。译码过程可以概括如下：

第 1 步：在译码之前，为每 2^{n-k} 个可能的伴随式 $\boldsymbol{s}$ 保存相应的最小加权矢量 $\boldsymbol{e}$，$\boldsymbol{e}$ 满足公式 $\boldsymbol{e}\boldsymbol{H}^{\mathrm{T}}=\boldsymbol{s}$；这样，就可以得到一个包含 2^{n-k} 个 n bit 分量的表格。

第 2 步：根据接收到的 n 维矢量 $\boldsymbol{y}$，通过线性运算 $\boldsymbol{s}=\boldsymbol{y}\boldsymbol{H}^{\mathrm{T}}$ 生成一个 $(n-k)$ 维的伴随式 $\boldsymbol{s}$；该过程需要一个 n 阶寄存器和 $n-k$ 个模 2 加法器。

第 3 步：对第 1 步得到的表格进行查询，以得到与第 2 步中伴随式 $\boldsymbol{s}$ 对应的 $\hat{\boldsymbol{e}}=\boldsymbol{e}$。

第 4 步：通过 $\boldsymbol{x}_m=\boldsymbol{y}\oplus\hat{\boldsymbol{e}}$ 运算来获取最相似的码矢量；该码矢量中，前 k 个符号就是数据符号。

2. 线性码的范例：Reed-Solomon 码

在发现了汉明码之后，循环码理论的发展诞生了第一批高效率的纠错码（BCH 码）系列。循环码的核心概念基础是有限域理论中的数学抽象概念，称为“理想概念”。循环码在实际应用中非常重要，因为它可以通过基于高速移位寄存器的编译码器来实现，数据传输速率可高达数 Gbit/s。在循环码系列中，有很多特殊而功能非常强大的纠错码，如 Golay 码、BCH 码、Reed-Solomon 码。循环码是根据生成多项式来分类的，生成多项式乘以数据多项式就可以得到码字多项式。这些多项式是在有限域中进行定义的，移位寄存器电路用来实现所需的多项式乘法和除法运算。目前比较流行的循环码称为“循环冗余校验（Cyclic Redundancy Check，CRC）码”，CRC 码是通过精简（减小块长度）某些循环码来得到的。CRC 码主要用于检错，也就是在接收到的消息中找出误码的位置。检错通常比纠错容易实现，主要用于对消息正确性要求很高的应用中；或者根据自动请求重发（ARQ）中规定的协议来标记出不可靠的消息位并丢弃该消息，同时请求重发。Reed-Muller 码（1954）已经不再使用了，但它是第一个可以纠正多个错误的纠错码；Reed-Muller 码具有很快的译码算法，但它们的数据传输速率太低。不过，Reed-Muller 码是构建现代复杂纠错码非常重要的部分。

Reed-Solomon（RS）码可以看做是 BCH 码的非二进制扩展，是一种非常有趣和实用的线性块状码。RS 码的块长度 n 为 $q-1$，其中 q 是指符号的字母表大小。我们可以发现，这些块状码只适用于大型字母表。包含 k 个信息符号、块长度为 n 的 RS 码的最短距离为 $d=n-k+1$，这个特性是“最大距离可分码”中所有纠错码共有的。RS 码非常适合产生成串误码的信道，这样的信道称为“突发”信道。如果符号的 $q=2^m$，那么块长度为 $n=2^m-1$。对于任意选择的奇数最短距离 d 来说，信息符号的数量为 $k=n-d+1$，可以纠正 $t=(d-1)/2=(n-k)/2$ 个错误。如果我们用 m 个二进制数字来表示码字中的每个字母，那么我们

就可以得到一个包含 km bit 的二进制码，其块长度为 nm bit。n 个二进制 m 元符号中最多包含 t 个误码的任何干扰序列都可以得到校正，因此，RS 码就可以校正所有长度不超过 $m(t-1)+1$ 的突发数据。

26.6 卷积码

卷积码可以由一个线性移位寄存器电路生成，该电路可以对信息序列进行卷积运算。在 20 世纪 50 年代末，卷积码可以通过时序译码算法来成功译码。1967 年，人们开发出了更简单的译码算法——Viterbi 算法，该算法使简单卷积码变得非常流行。但是，Viterbi 算法在更强大的卷积编码中是不实用的。

一个速率为 $R=1/n$ 的卷积编码器是一个线性有限状态机，该状态机包含 1 个输入端、n 个输出端和 1 个 K 阶移位寄存器；其中 $K=m+1$ 称为卷积码的"约束长度"，m 是指有限状态机的存储器数量。图 26-4 给出了一个编码器的示例，其中 $R=1/2$，$K=3$。这样的编码器具有 2^m 个可能的状态，可以通过移位寄存器的各个支路组合来表示；这些移位寄存器通过模 2 加法运算来形成编码器的输出结果。每个输出端可以通过一个带有二进制系数的多项式来描述；其中，系数 1 表示一个支路，系数 0 表示没有支路。图 26-4 中的例子描述了每个输入位如何生成两个输出符号的过程，该输出符号是通过一个当前的输入位和两个先前的输入位的卷积运算得到的，这两个先前的输入位保存在图中阴影部分的寄存器单元中。

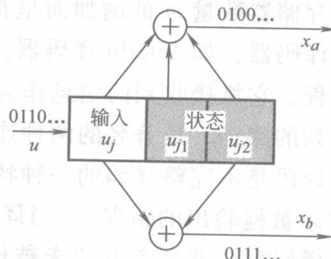


图 26-4 卷积编码示例（$K=3$）

图 26-4 中的有限状态机可以通过一个状态图来描述，如图 26-5a 所示。图中的状态转移是通过定向边来描述的，该定向边上有一个标签 $u_j/x_{0j}\ x_{1j}$；该标签的含义是指状态从 $u_{j-1}u_{j-2}$ 转移到 u_ju_{j-1} 时与之相关的一个输入和两个输出。图 26-5b 中的栅格图等效描述了该有限状态机的状态转移过程，该图着重从时间的角度描述了各个状态之间的变化。该图中从左至右的路径对应了编码器的状态变化。类似于块状码，卷积码的特性也依赖于各个输出符号流之间的汉明距离，这些汉明距离对应了纠错码的各个码字。性能强大的卷积码具有很多最短距离，这些最短距离可以通过查找各个路径的最小输出加权值（1 的数量）来确定，这些路径都从状态 00 出发，经过不同的中转之后，到达相同的状态终点。由于卷积码的线性原因，选择哪一个参考状态是无关紧要的。图 26-5b 中的栅格图是实现高效率译码方法——Viterbi 算法（1967）的关键，该算法是根据发明者 Viterbi 的名字命名的。Viterbi 算法是一种高效率的译码方法，该算法将接收符号流的距离与栅格图中所

有可能的路径进行了比较，从中选出最短的汉明距离；该算法的高效之处在于，在每个状态处（栅格图中的各个节点）路径的累积距离都可以进行比较，而且只保留距离最短的路径，并将其他路径永远丢弃，同时不会影响路径的最优性；这一点大大降低了计算路径时的负载。一旦这个择优过程完成，保留下来的路径就是接收序列最短路径的候选路径。如果栅格图在某一时刻被确定下来，那么最后就只有一个路径是到达状态00的最佳路径。一旦这个最佳路径被选定，译码位的对应序列只能沿着最佳路径的反向路径进行读取。事实上，Forney在1973年就发现，Viterbi算法实现了最佳似然译码过程，即最佳译码策略。

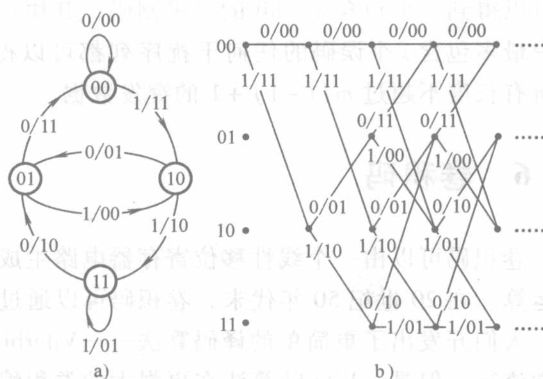


图26-5　状态转移和栅格图

连续译码是另一种更早用来对卷积码进行译码的方法，它的复杂性不会随着编码存储器数量m的增加而呈指数增长，但它的性能只接近于一个真正的最大似然译码器，如Viterbi译码器。连续译码是一个通过“代码树”进行系统搜索的过程，它将接收到的信息作为向导，其目标就是最后追踪到代表着实际发送信息序列的路径。最著名的两种连续译码算法就是“栈式算法”和“Fano算法”。连续译码是不完整译码的一种特例，不完整译码会产生两种译码故障。第一种称为“未被检验出的错误”，当译码器接收了很多错误的假定条件并在这些条件下进行译码时，就会产生“未被检验出的错误”。第二种称为“缓冲溢出”，当每个数据块的计算量都超过允许值时，就会发生缓冲溢出；在这种情况下，数据帧就无法进行译码，并称为“删除帧”。卷积码也是连续译码，它具有足够大的约束长度，以便与数据块在允许时间内无法成功译码的概率相比时，译码器的“未被检验出的错误”概率可以忽略不计。

在对编码增益要求很高和误码概率要求很低的应用中，广泛采用了层叠两个代码的方法，称为“链接码”。功能最强大的组合译码方法就是外部采用Reed-Solomon码，内部采用卷积码；该译码过程分为两个阶段，这使得该译码过程变成了次优方案，但它的性能非常接近于最佳译码器。由于内部译码器会产生突发误码，因此通常需要使用一种称为“交叉”的技术来分散这些误码并减轻外部译码器的负载。通常，内部Viterbi译码器会校正这些误码，以便高速的外部编码可以将误码率降低到理想的程度。

Turbo 码

编码理论创立者通常处理的问题是如何通过开发具有很多结构的代码来构建功能强大的编码方案，这就是可行性译码器产生的原因。但是，编码理论指出如果数据块的长度足够大的话，随机选择的编码方案应该是性能非常好的方案。但是直到现在，人们还没有意识到为所有随机的大型代码寻找可行性译码器所面临的挑战性。

1993 年，诞生了一种新的链接码，称为“Turbo 码”。Turbo 码既可以看做是卷积码，也可以看做是块状码；Turbo 码可以在一定的译码复杂性条件下实现接近 Shannon 极限的纠错性能。

Turbo 编码器是两个简单卷积编码器的组合体。对于一个包含 k 个信息位的数据块，每个代码都会产生一组奇偶位。Turbo 码是由信息位和奇偶位构成的，如图 26-6 所示。Turbo 编译码器中的关键创新之处就是数字复用器 P，数字复用器在对下一个代码进行编码之前，会改变原始 k 个信息位的次序。如果数字复用器选择恰当，那么代码中容易出错的码字对应的信息块将与其他代码中不会出错的码字对应起来；最后得到的代码的性能与 Shannon 随机码非常相似，Shannon 随机码可以实现最佳的译码性能，不过其译码器比较复杂。Turbo 译码过程采用了两个与各代码匹配的简单译码器；每个译码器会向另一个译码器发送似然性估计，并将从另一个译码器发送来的对应估计作为“先验似然性”。Turbo 译码器在两个子译码器的输出信号之间进行迭代，直至达到满意的收敛性。

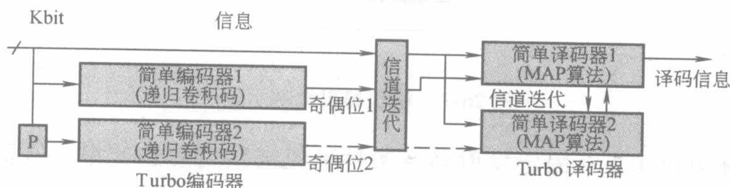


图 26-6　Turbo 编译码器

Turbo 码的性能超过了目前所知的功能最强大的编码方案，更重要的是 Turbo 码的译码更加简单。为了实现这么卓越的性能，Turbo 码需要使用大型数字复用器，但该数字复用器没有目前链接码中使用的数字复用器大。

26.7　格状编码调制

为了提高相加性噪声和码间干扰信道（带宽严重受限的信道，如电话信道）上的信噪比，人们提出了最原始的“分组”技术，该技术最典型的例子就是 Ungerboeck 在 1982 年开发的格状编码调制（Trellis Coded Modulation，TCM）技术。

Ungerboeck给出了在编码信号与调制信号之间的对应关系没有清楚确定的前提下，不同数据序列之间的汉明距离与调制数据序列之间的欧几里德（Euclidean）距离之间的关系。他提出了一种新的分组方法，该方法不仅可以为编码信号分配信号，还可以为调制数据序列之间的欧几里德距离下限提供简单的计算公式。图26-7b和26-7c给出了编码器和格状编码调制的调制信号分配示例，该示例将与图26-7a中未编码的四相相移键控（Quadri-Phase Shift Keying，QPSK）方案进行比较。

格状编码（见图26-7b）用一个8进制的PSK信号星座取代了QPSK信号星座，从而在复平面上形成了8个点，该复平面代表了相位调制过程。格状编码每次不是将8个点中的3个点调制成一个信道符号，而是每次通过将2个数据位调制成一个3bit的信道符号来提高数据传输速率的。这3bit定义了图26-7c相位图中标示的1/8概率。每一对输入数据位都会被格状编码器映射成3个代码，从而产生一个波形。发送的格状编码波形具有和QPSK波形相同的数据传输速率，并使用相同的带宽，但其功率要求降低了2.5倍。

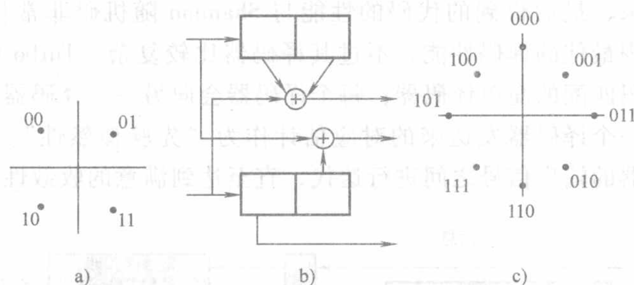


图26-7 格状编码调制示例

TCM还可以用来在固定发射功率条件下提高数据发送速率；但要实现这个目标就需要更多复杂的信号星座。

26.8 应用

纠错码已经广泛应用到了各种通信系统中。从计算机终端、飞机到太空飞船，数字数据已经无处不在。即使在接收信号的功率非常接近热噪声功率的情况下，信号编码都可以用来实现可靠的通信连接。随着无线频谱空间变得越来越拥挤，纠错编码也变得越来越重要了，因为纠错编码可以为处于干扰环境中的通信连接提供可靠性保证。这一点在军事应用中尤其重要，因为在军事应用中通常可以采用纠错编码来抵抗敌方蓄意的干扰。

纠错编码也是通信系统中用来降低功率需求的一种很好的方法，尤其是在功

率比较缺乏的系统中，如通信中继卫星中，因为功率较弱的消息可以在编码过程中得到正确的恢复。

在计算机系统的数据流中，即使很低的误码率也不能容忍；因为单个误码就可以破坏一个计算机程序，因此在这些应用中，纠错码就变得越来越重要了。数据流可以通过简单汉明码来进行分组编码并保存到密集的计算机存储器中。目前，各种编码方案已经广泛用于保护各种数字磁带和磁盘中的数据了。

复杂系统中的通信同样非常重要，而通常在各个子系统之间都存在很大的数据流量，如在数字交换系统和数字雷达信号处理系统中。这些内部数据既可以通过电路线进行传输，也可以通过更加复杂的时分数据总线系统来进行传输。无论哪一种情况，纠错技术都可以用来保证可靠的传输性能。

名词解释

块状码：一种编码方案；选择长度为 k 的信息块，产生一个长度为 n 的码字（$n>k$）（n 是指信息块长度或码长）。

突发纠错：一种纠错方法；可以有效抵抗成串的误码出现，但无法纠正随机误码。

信道容量：在最低误码率的条件下，信道上可以传输的最大信息速率。

编码率：信息块中信息位数 k 与码字中发送符号总数的比值（$n=k+$冗余符号数）。

码字：通过在原始消息中添加冗余符号得到的符号序列。

链接码：通过层叠实现的编码；链接码通过数字复用器来分离。

卷积码：由线性移位寄存器电路产生的编码；线性移位寄存器电路在信息序列中执行卷积运算。

循环码：根据生成多项式定义的编码；该生成多项式乘以数据多项式后就可以得到码字多项式。这些多项式都是定义在有限域中，移位寄存器电路用来执行所需的多项式乘法和除法运算。典型的循环码包括：Golay 码、BCH 码和 Reed-Solomon 码。

纠错：从码字中检测误码的一种机制；无法检测出来的误码模式称为“未被检验出的错误”。

Galois 域：Galois 域（有限域）是一组对象或单元，+和·运算都定义在这些对象或单元上，并遵守某些特性的规则（例如，运算具有交换性和分配性）。例如：如果 p 是素数，那么整数 $\{0, 1, \cdots, p-1\}$ 以 p 为模就可以形成一个有限域。

生成矩阵：一个 $k\times n$ 阶矩阵；通过与信息块（长度为 k）相乘就可以生成

一个块状码码字（长度为 n）。

汉明距离：两个具有相同长度的序列之间不同位的数量；一个代码的最短距离是指任何一对码字之间的最小汉明距离。

数字复用器：一种用来打乱码字符号顺序的设备；这样给定码字的符号之间就可以充分相互隔离。

线性码：一种编码；其中任何码字的组合仍然是一个码字。

最大似然译码：一种译码方法；该过程在给定任何可能码字的条件下，使接收序列的概率最大化。如果所有码字的概率都相同，这种译码方法就会产生最小的误码概率。

冗余：添加在原始消息中的附加性符号（奇偶校验符号），用于纠错。

连续译码：一种通过代码树进行系统搜索的译码方法，其目标是得到最接近接收序列的路径方案。典型的连续译码算法包括：Fano 算法、栈式算法。

伴随式：通过译码器计算得到的一个矢量，用来查找可校正误码的位置。

系统码：一种编码，其中每个码字的前 k 个符号与信息块的完全相同。所有的编码都可以转换成系统码的形式。

格状编码调制：一种联合编码和调制的组合方法，其基础是卷积码的设计必须与调制信号组相匹配，这样可以使调制序列之间的欧几里德距离最大。

栅格图：用来描述有限状态机实时更新情况的图。在该图中，各个状态被定义成各个目的顶点，而各种可能的状态转移被定义成各个边。

Turbo 码：由两个（或更多个）并联的简单卷积编码器生成的一种编码；并联的卷积译码器通过数字复用器来进行分离。

Viterbi 算法：一种高效率的卷积译码方法，其原理是在栅格图中寻找接收序列的最短距离路径。

加权值分布：关于每个码字与给定参考码字之间的汉明距离的列表。

参考文献

[1] Berlekamp. E. R. 1968. *Algebraic Coding Theory*. McGraw-Hill, New-York.

[2] Berlekamp. R. R. , ed. 1974. *Key Papers in the Development of Coding Theory*. IEEE Press, New-York.

[3] Blahut, R. 1983. *Theory and Practice of Error Control Codes*. Addison-Wesley. Reading, MA.

[4] Bose, R. C. and Ray-Chaudhuri, D. K. 1960. On a class of error correcting binary group codes. *Info. And Control* 3: 68 ~ 79.

[5] Clark, G. C. and Cain, J. B. 1981. *Error-Correction Coding for Digital Communication*. Plenum, New York.

[6] Fire, P. 1959. A class of Multiple-Error-Correcting Binary Codes for Non-independent Errors. Sylvania Report No. RSL-E-2, Sylvania Electronic Defense Laboratory, Reconnaissance Systems Division, Mountain View, Calif. , March.

[7] Forney, G. D. 1973. *The Viterbi algorithm*. Proc. IEEE 61: 268 ~ 278.

[8] Forney, G. D. 1967. *Concatenated Codes*. MIT Press, Cambridge, MA.

[9] Hamming, R. W. 1980. *Coding and Information Theory*. Prentice-Hall, Englewood Cliffs. NJ.

[10] Hocquenghem, A. 1959. Codes correcteurs d'erreurs. *Chiffres* 2: 147 ~ 156.

[11] Lidl, R. and Niederreiter, H. 1983. *Finite Fields*. Addison-Wesley, Reading, MA.

[12] MacWilliams, F. J. and Sloane, N. J. 1977. *The Theory of Error-Correcting Codes*. North-Holland, Amsterdam.

[13] Muller. D. E. 1954. Application of Boolean algebra to switching circuit design and to error detection. *IEEE Trans. Computers* 3: 6 ~ 12.

[14] Reed, I. S. 1954. A class of multiple-error-correcting codes and the decoding scheme. *IEEE Trans. Info. Theory* 4: 38 ~ 49.

[15] Reed, I. S. and Solomon, G. 1960. Polynomial codes over certain finite fields. *J. SIAM* 8: 300 ~ 304.

[16] Schouhamer Immink, K. A. 1989. Coding techniques for the noisy magnetic recording channel—A state-of-the-art report. *IEEE Trans. Comm.* 37 (5): 413 ~ 419.

[17] Schouhamer Immink, K. A. 1991. *Coding techniques for Digital Recorders*. Prentice-Hall, Englewood Cliffs, NJ.

[18] Shannon, C. E. 1948. A mathematical theory of communication. *Bell Sys. Tech. J.* 27: 379: 423, 623: 656.

[19] Ungerboeck, G. 1982. Channel coding with multilevel/phase signals. *IEEE Trans. Info. Theory* (Jan.).

[20] Viterbi, A. J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* IT-13: 260 ~ 269.

[21] Viterbi, A. J. and Omura, J. K. 1979. *Principles of Digital Communication and Coding*. McGraw-Hill, New-York.

备注

本章中的大多数编码系统如果进行完整详细的讨论的话，都过于复杂。如果读者想更进一步了解编码理论中的各种概念，建议可以参考 Clark 与 Cain（1981）、Berlekamp（1974）、Berlekamp（1968）、MacWilliams 与 Sloane（1977）、Blahut（1983）、Viterbi 与 Omura（1979）以及 Hamming（1980）的著作。这些参考著作只是介绍纠错码技术的一部分资料。

关于磁性记录编码技术的主题，读者可以参考 Schouhanmer Immink 在 1989 年的著作；而关于压缩磁盘编码的问题，读者可以参考 Schouhanmer Immink 在 1991 年的著作；关于链接码的主题，读者可以参考 Forney 在 1967 年的著作；关于有限域的重要数学工具，读者可以参考 Lidl 和 Niederreiter 在 1983 年的著作。

缩 略 语

缩写词	英 文 全 称	中 文 全 称
ABR	Associativity-Based Routing	基于关联性的路由算法
ACL	Asynchronous Connectionless Link	异步无连接
A/D	Analog-to-Digital	模/数转换
ADP	Accumulative Difference Pictures	累积差分图
ADPCM	Adaptive Differential Pulse Code Modulation	自适应差分脉冲编码调制
AI	Artificial Intelligence	人工智能
AIFS	Arbitration InterFrame Space	判优帧间空间
AODV	Ad Hoc On demand Distance Vector	Ad Hoc 需求距离矢量算法
AS	Advanced Schottky	先进肖特基
ALS	Advanced Low-power Schottky	先进低功耗肖特基
ALU	Arithmetic Logical Unit	算术逻辑单元
AMA	Active Member Address	活动成员地址
ARA	Ant-colony-based Routing Algorithm	基于"蚁群"的路由算法
ARQ	Automatic Repeat reQuest	自动重发请求
ASIC	Application Specific Integrated Circuits	专用集成电路
ASSP	Application Specific Standard Product	标准专用产品
ATCP	Ad Hoc TCP	Ad Hoc TCP 层
ATE	Automated Test Equipment	自动测试设备
ATM	Asynchronous Transfer Mode	异步传输模式
BAM	Bidirectional Associative Memory	双向相联存储器
BE	Best Effort	尽力服务
BiCMOS	Bipolar CMOS	双极性 CMOS
BIST	BuiltIn-Self-Test	内置自行测试
BGA	Ball Grid Array	球状栅格阵列
BPF	Band-Pass Filter	带通滤波器
BRDF	Bidirectional Reflectance Distribution Function	双向反射分布函数
BSF	Band-Stop Filter	带阻滤波器
BTr	receive Busy Tone	接收忙音
BTt	transmit Busy Tone	发送忙音
CAD	Computer Aided Design	计算机辅助设计
CAD/CAM	Computer Aided Design/Computer Aided Manufacturing	计算机辅助设计/计算机辅助制造

（续）

缩写词	英文全称	中文全称
CAM	Content Addressable Memory	内容可编址存储器
CAS	Column Address Strobe	列地址选通
CASE	Computer Aided Software Engineering	计算机辅助软件工程
CAV	Constant Angular Velocity	恒定角速度
CBRP	Cluster-based Routing Protocol	基于簇的路由协议
CCD	Charge Coupled Device	电荷耦合器件
CDMA	Code Division Multiple Access	码分多址
CD-R	Compact Disc-Recordable	可记录压缩盘
CGSR	Clusterhead-Gateway Switch Routing	簇头网关交换路由算法
CICC	Custom Integrated Circuits Conference	特定集成电路会议
CISC	Complex Instruction Set Computer	复杂指令集计算机
CLM	Complex Logarithmic Mapping	复杂对数映射
CLV	Constant Line Velocity	恒定线速度
CMIP	Common Management Information Protocol	公共管理信息协议
CMM	Color Management Methods	颜色管理方法
CMOS	Complementary Metal Oxide Semiconductor Transistor	互补金属氧化物半导体
CMOT	CMIP over TCP/IP	TCP/IP 网上的公共管理信息协议
CPI	Cycles Per Instruction	每条指令占用的时钟周期
CPLD	Complex Programmable Logic Device	复杂可编程逻辑器件
CPU	Central Processing Unit	中央处理单元
CRC	Cyclic Redundancy Check	循环冗余校验码
CRT	Cathode Ray Tube	阴极射线管
CSMA/CD	Carrier sense multiple Access with Collision Detection	带有检测冲突的载波侦听多路存取协议
CSG	Constructive Solid Geometry	结构立体几何
CT	Computerized Tomography	计算机断层造影
CV	Constant Voltage	常压
CVGIP	Computer Vision, Graphics, and Image Processing	计算机视觉、图形和图像处理
CVPR	Conference on Computer Vision and Pattern Recognition	计算机视觉和模式识别会议
CW	Contention Window	竞争窗口
D/A	Digital-to-Analog	数/模转换

（续）

缩写词	英文全称	中文全称
DAC	Design Automation Conference	设计自动化会议
DBMS	Data Base Management System	数据库管理系统
DBTF	Development Before The Fact	事先预防
DBTMA	Dual Busy Tone Multiple Access	双忙音多址协议
DCE	Data Communications Equipment	数据通信设备
DCF	Distributed Coordination Function	分布式协同功能
DCFL	Direct Coupled FET Logic	直接耦合 FET 逻辑
DDR	Distributed Dynamic Routing	分布式动态路由
DESC	Defense Electric Supply Center	(美国)国防部电子支持中心
DFT	Design For Test	测试设计
DIE	Die Information Exchange	冲模信息交换
DIFS	DCF Interframe Space	DCF 帧间空间
DLL	Delay-Locked Loop	延迟锁环
DPA	Destructive Physical Analysis	破坏性物理分析
dpi	dots per inch	每英寸上的墨点数
DOD	Drop On Demand	按需喷墨技术
DPM	Defects Per Million	每百万个器件中存在缺陷的数量
DRAM	Dynamic Random Access Memory	动态随机存取存储器
DRAM	Direct Random Access Memory	直接随机存取存储器
DREAM	Distance Routing Effect Algorithm for Mobility	移动远距离路由效应算法
DSDV	Destination-Sequential Distance-Vector Routing	目的地连续远距离矢量路由
DSE	Data Switching Exchanges	数据交换机
DSP	Digital Signal Processing	数字信号处理
DSR	Dynamic source Routing	动态源节点路由
DST	Distributed Spanning Tree	基于分布式生成树的路由
DTC	Diode-Transistor Logic	二极管-晶体管逻辑
DTE	Data Terminal Equipment	数据终端设备
DUT	Device under test	待测器件
DV	Distant Vector	远距离矢量
EAROM	Electrical Alterable ROM	电改写只读存储器
ECC	Error Correction Code	纠错码
ECCV	European Conference on Computer Vision	计算机视觉欧洲会议

（续）

缩写词	英文全称	中文全称
ECL	Emitter Couple Logic	发射极耦合逻辑
ECN	Explicit Congestion Notification	精确拥塞通知
EDA	Electronic Design Automation	电子设计自动化
EDCA	Enhanced Distributed Channel Access	增强分布式信道接入
EDO	Extended Data Output	扩展数据输出
E/D FET	Enhancement and Depletion FET	增强型和耗尽型 FET
EEPROM	Electrical Erasable Programmable Read-Only Memory	电可擦可编程只读存储器
EM	Expectation-Maximization	最大期望值
EMI	Electro Magnetic Interference	电磁干扰
EMP	Ego-Motion Polar	自我运动极性
EY-NPA	Elimination-Yield Nonpreemptive Priority Access	非优先排除接入法
EPROM	Erasable Programmable Read Only Memory	可擦除可编程只读存储器
ESD	ElectroStatic Discharge	静电放电
FAST	Fairchild Advanced Schottky TTL	Fairchild 先进肖特基 TTL
FC-AL	Fiber Channel-Arbitrated Loop	光纤判决信道环
FDDI	Fiber Distributed Data Interface	光纤分布式数据接口
FDMA	Frequency Division Multiple Address	频分多址
FEC	Forward Error Correction	前向纠错
FES	Focus Error Signal	聚焦误差信号
FFT	Fast Fourier Transform	快速傅里叶变换
FHSS	Frequency Hopping Spread Spectrum	跳频扩频
FIFO	First In First Out	先进先出
FIR	Finite Impulse Response	有限脉冲响应
FIT	Failure In Time	故障率
FMR	Focus MisRegisteration	聚焦重合失调
FOE	Focus Of Expansion	伸展中心
FORP	Flow Oriented Routing Protocol	面向数据流的路由协议
FPGA	Field Programmable Gate Arrays	现场可编程门阵列
FPM	Fast Page Mode	快速页面模式
FPU	Float Point math Unit	浮点数学单元
FR	Free for Receive	自由接收

（续）

缩写词	英文全称	中文全称
FSLS	Fuzzy Sighted Link State	模糊可见链接状态算法
FSR	Fisheye State Routing	鱼眼状态路由算法
FT	Free for Transmission	自由发送
GRNN	General Regression Neural Network	广义回归神经网络
GSR	Global State Routing	整体状态路由算法
GUI	Graphical User Interfaces	图形用户界面
HDL	Hardware Description Language	硬件描述语言
HFET	Heterostructure Field Effect Transistor	异质结构场效应晶体管
HMP	Hidden Markov Process	隐蔽马尔可夫过程
HPF	High Pass Filter	高通滤波器
HPM	Hyper Page Mode	超页面模式
HSR	Hierarchical State Routing	分级状态路由算法
IAPR	International Association for Pattern Recognition	国际模式识别联盟
IC	Integrated Circuit	集成电路
ICASSP	International Conference on Acoustics, Speech and Signal Processing	国际声学、语音和信号处理会议
ICC	International Color Consortium	国际色彩联盟
ICCV	International Conference on Computer Vision	国际计算机视觉会议
ICDAR	International Conference on Document Analysis and Recognition	国际文档分析与识别会议
ICMP	Internet Control Message Protocol	互联网控制消息协议
ICPR	International Conference on Pattern Recognition	国际模式识别会议
IEEE	Institute of Electrical and Electronics Engineers	电气和电子工程师学会[美]
ISM	Industrial Scientific and Medical	工业、科学和医学
ITC	International Test Conference	国际测试会议
I^2L	Integrated Injection Logic	集成注入逻辑
IIR	Infinite Impulse Response	无限脉冲响应
I/O	In/Out	输入/输出
ISL	Integrated Schottky Logic	集成肖特基逻辑
ISO	International Organization for Standardization	国际标准化组织
ISSCC	International Solid State Circuits Conference	国际固态电路会议
IS&T	Society for Imaging Science and Technology	成像科学和技术协会

（续）

缩写词	英 文 全 称	中 文 全 称
IWGR	International Workshop on Graphics Recognition	国际图形识别研习会
JAN	Joint Army Navy	联合陆军海军
JEDEC	Joint Electron Device Engineering Council	电子工程设计发展联合会议
JFET	Junction Field Effect Transistor	结型场效应晶体管
KGD	Known Good Die	知名优质冲模
LA-MAC	Load Awareness MAC	负载认知 MAC 协议
LAN	Local Area Network	局域网
LAR	Location Aided Routing	位置辅助路由算法
LET	Link Expiration Time	链接截止时间
LFSR	Linear Feedback Shift Register	线性反馈移位寄存器
LMR	Light-weight Mobile Routing	轻型移动路由算法
LMS	Least Mean Square	最小均方
LORA	Least Overhead Routing Approach	最低开销路由法
LPF	Low-Pass Filter	低通滤波器
LSB	Least Significant Bit	最低有效位
LS	Link-State	链接状态
LSI	Large Scale Integration	大规模集成电路
LTI	Linear Time Invariant	线性时不变
MAC	Medium Access Control	媒体接入控制
MACA-BI	Multiple Access with Collision Avoidance By Invitation	基于邀请的冲突避免多址接入
MAC-RSV	MAC-ReSerVation	MAC 预留
MAN	Metropolitan Area Network	城域网
MAU	Multistation Access Unit	多站接入单元
MCM	MultiChip Module	多片组件
MCMC	Multichip Module Conference	多片组件会议
MCMO	Moving Camera/Moving Objects	移动摄像机/移动目标
MCM-C	Ceramic MultiChip Module	陶瓷多片组件
MCM-D	Deposited MultiChip Module	沉积多片组件
MCM-L	Laminate MultiChip Module	碾压多片组件
MCSO	Moving Camera/Stationary Objects	移动摄像机/固定目标
MESFET	MEtal Semiconductor Field Effect Transistor	金属半导体场效应晶体管

（续）

缩写词	英 文 全 称	中 文 全 称
MIMD	Multiple Instruction Multiple Data	多指令多数据
MISD	Multiple Instruction Single Data	多指令单数据
MISR	Multiple Input Shift Register	多输入移位寄存器
ML	Maximum Likelihood	极大似然
MMU	Memory Management Unit	存储器管理单元
MMWN	Multimedia support in mobile Wireless Networks	移动无线网络中的多媒体支持
MNOS	Metal Nitrite Oxide Silicon	金属亚硝酸氧化物硅
MO	Magneto-Optic	磁-光技术
MODFET	MOdulation Doped Field Effect Transistor	调节掺杂场效应晶体管
MOSFET	Metal-Oxide-Semiconductor Field Effect Transistor	金属-氧化物-半导体型场效应晶体管
MOSIS	MOS Implementation System	MOS 执行系统
MPR	MultiPoint Relays	多点延时
MRI	Magnetic Resonance Imaging	核磁共振成像
MRL	Message Retransmission List	消息重发目录
MSB	Most Significant Bit	最高有效位
MSDU	MAC Service Data Unit	MAC 业务数据单元
MSE	Mean Square Error	均方误差
MSI	Medium-Scale Integration	中规模集成电路
NA	Numerical Aperture	数值孔径
NAV	Network Allocation Vector	网络分配矢量
NIC	Network Interface Card	网络接口卡
NOS	Network Operating System	网络操作系统
NRE	Nonrecurring Engineering	一次性工程
NRL	Normalized Residual Lifetime	标准剩余生命周期
OA	Organic Acid	有机酸
OCR	Optical Character Recognition	光学字符识别
OEIC	OptoElectronic Integrated Circuits	光电子集成电路
OMPAC	OverMolded Pad-Array Carrier	过模片状阵列载体
OLSR	Optimized Link State Routing	最佳链接状态路由算法
OPC	Organic PhotoConductor	有机光电导体
OSI	Open System Interconnect Reference Model	开放式系统互联参考模型

（续）

缩写词	英 文 全 称	中 文 全 称
PAD	Packet Assembler Disassembler	分组组合器/拆分器
PAL	Programmable Arrays Logic	可编程阵列逻辑
PC	PhotoConductor	光电导体
PCB	Printed Circuit Board	印制电路板
PCDC	Power Controlled Dual Channel	功耗控制双信道协议
PCF	Point Coordination Function	点协同功能
PCM	Pulse Code Modulation	脉冲编码调制
PCS	Profile Connection Space	规则映射空间
PDL	Page Description Language	页面描述语言
PDN	Public Data Networks	公共数据网
PDU	Protocol Data Unit	协议数据单元
PGA	Pin Grid Array	引脚网格阵列
PHY	PHYsical	物理层
PIND	Particle Impact Noise Detection	微粒影响噪声探测
PLA	Programmable Logic Arrays	可编程逻辑阵列
PLD	Programmable Logic Device	可编程逻辑器件
PLL	Phase-Locked Loop	锁相环
PMA	Parked Member Address	停机成员地址
PNN	Probabilistic Neural Network	随机性神经网络
ppm	part per million	百万分之…
PPM	Pulse Position Modulation	脉冲位置调制
PQFP	Plastic Quad FlatPack	塑封方块扁平封装
PRMA	Packet Reservation Multiple Access	数据包预留多址
PRML	Partial Response Maximum Likelihood	部分响应最大相似性
PROM	Programmable Read-Only Memory	可编程只读存储器
PSE	Packet Switching Exchanges	分组交换机
PWM	Pulse Width Modulation	脉冲宽度调制
QFP	Quad FlatPack	方块扁平封装
QoS	Quality of Service	服务质量
QPSK	Quadri-Phase Shift Keying	四相相移键控
RAC	Reliability Analysis Center	可靠性分析中心
RAID	Redundant Arrays of Independent Disk	独立磁盘冗余阵列

（续）

缩写词	英文全称	中文全称
RAM	Random Access Memory	随机存取存储器
RAS	Row Address Strobe	行地址选通
RB	Receiver Beacon	接收端信标
RC	Resistance-Capacitance	电阻-电容
RCTL	Resistor-Capacitor-Transistor Logic	电阻-电容-晶体管逻辑
RDMAR	Relative Distance Micro-discovery Ad Hoc Routing	相对距离微发现 Ad Hoc 路由算法
RE/TM	Rare Earth/Transition Metal	稀有地球/过渡金属
RF	Radio Frequency	射频
RFI	Radio Frequency Interference	无线电频率干扰
RFN	Route Failure Notification	路由故障通知
RIMA-DP	Receive Initiated Multiple Access with Dual-purpose Polling	接收端启动双重目的轮询多址
RIMA-SP	Receive Initiated Multiple Access with Simple Polling	接收端启动单轮询多址
RISC	Reduced Instruction Set Computer	精简指令集计算机
RPM	Revolution Per Minute	每分钟转速
ROAM	Routing On-demand Acyclic Multi-path	路由需求非循环多路径算法
ROM	Read Only Memory	只读存储器
RR	Reserved for Receive	预留接收
RRN	Route Reestablishment Notification	路由重建通知
RS	Reed-Solomon	里德-所罗门码
RT	Reserved for Transmission	预留发送
RTL	Resistor-Transistor Logic	电阻-晶体管逻辑
RTO	Retransmit TimeOut	重发定时
RTR	Ready-To-Receive	准备接收
RTS/CTS	Request-To-Send/Clear-To-Send	请求发送/清除发送
RwoH	Reliability without Hermeticity	无密封可靠性
SAR	Successive Approximation Register	逐次近似寄存器
SCA	Single Connector Attachment	单连接器附件
SCFT	Source Coupled FET Logic	源极耦合 FET 逻辑
SCK	Scan-Clock	扫描时钟
SCMO	Stationary Camera/Moving Objects	固定摄像机/移动目标
SCO	Synchronous Connection Oriented	同步面向连接

（续）

缩写词	英文全称	中文全称
SCR	Silicon Control Rectifier	硅控整流器
SCSI	Small Computer Systems Interface	小型计算机系统接口
SCSO	Stationary Camera/Stationary Objects	固定摄像机/固定目标
SDMA	Space Division Multiple Access	空分多址
SEI	Software Engineering Institute	软件工程协会
SI	Scan-In	扫描输入
SIFS	Short Interframe Space	短帧间空间
SIR	Signal-to-Interference Ratio	信干比
SISD	Single Instruction Single Data	单指令单数据
SIMD	Single Instruction Multiple Data	单指令多数据
SLURP	Scalable Location Update Routing Protocol	可升级位置更新路由协议
SMDS	Switched Multimegabit Data Service	交换式多兆位数据业务
SNMP	Simple Network Management Protocol	简单网络管理协议
SO	Scan-Out	扫描输出
SOO	System Oriented Object	面向系统对象
SPARC	Scalable Processor ARChitecture	可缩放处理机体系结构
SPIE	Society of Photo-Optical Instrumentation Engineers	光学照相仪器工程师协会
SRAM	Static Random Access Memory	静态随机存取存储器
SQL	Structured Query Language	结构化查询语言
SSA	Signal Stability Adaptive	信号稳定性适应算法
SSI	Small Scale Integration	小规模集成电路
STAR	Source-Tree Adaptive Routing	源树形适应性路由算法
STL	Schottky Transistor Logic	肖特基晶体管逻辑
STP	Shielded Twisted Pair	屏蔽双绞线
TAB	Tape Automated Bonding	带状自动压焊
TBRPF	Topology Broadcast based on Reverse Path Forwarding	基于反向路径的拓扑广播
TCM	Trellis Coded Modulation	格状编码调制
TDD	Time Divided Duplex	时分复用
TDMA	Time Division Multiple Access	时分多址接入
TES	Tracking Error Signal	寻轨误差信号
TMR	Tracking MisRegisteration	寻轨重合失调

（续）

缩写词	英文全称	中文全称
TOC	Table Of Contents	目录表
TORA	Temporarily Ordered Routing Algorithm	临时整齐路由算法
TTL	Transistor-Transistor Logic	晶体管-晶体管逻辑电路
TxOP	Transmission Opportunity	传输机会
ULSI	Ultra Large Scale Integration	极大规模集成电路
UTP	Unshielded Twisted Paire	非屏蔽双绞线
VAN	Value Added Network	增值网
VCO	Voltage Control Oscillator	电压控制振荡器
VIMS	Visual Information Management System	视觉信息管理系统
VLIW	Very Long Instruction Word	超长指令字
VLSI	Very Large Scale Integration	超大规模集成电路
VTC	Voltage Transfer Characteristic	电压转移特性
V&V	Verification and Validation	认证和审定
WAN	Wide Area Network	广域网
WEEE	Waste Electrical and Electronic Equipment	损耗电气和电子设备
WMN	Wireless Mesh Network	无线 Mesh 网络
WOM	Write Once Memory	一次性写入存储器
WRP	Wireless Routing Protocol	无线路由算法
WSI	Wafer Scale Integration	晶片规模集成
WTA	Winner Takes All	优胜劣汰
ZHLS	Zone-based Hierarchical Link State	基于区域的分级链接状态
ZRP	Zone Routing Protocol	区域路由协议

国际视野　科技前沿

国际信息工程先进技术译丛

《WCDWA原理与开发设计》

《下一代移动系统：3G/B3G》

《IMS:IP多媒体概念和服务》（原书第2版）

《下一代无线系统与网络》

《深入浅出UMTS无线网络建模、规划与自动优化：理论与实践》

《HSDPA/HSUPA技术与系统设计——第三代移动通信系统宽带无线接入》

《无线传感器及元器件：网络、设计与应用》

《印制电路板——设计、制造、装配与测试》

《IPTV与网络视频：拓展广播电视的应用范围》

《P2P系统及其应用》

《多电压CMOS电路设计》

《微电子技术原理、设计与应用》（原书第2版）



上架指导：工业技术/电子技术

CRC Press
Taylor & Francis Group

ISBN 978-7-111-23879-9



ISBN 978-7-111-23879-9

封面设计：马精明

编辑热线：(010)88379764

地　址：北京市百万庄大街22号　邮政编码：100037
联系电话：(010)68326294　网址：http://www.cmpbook.com(机工门户网)
(010)68993821　E-mail:cmp@cmpbook.com
购书热线：(010)88379639　(010)88379641　(010)88379643

定价：68.00元